Encoding EEG Signals to Examine Human-Like Next-Word Prediction Behaviour in Language Models

Boi Mai Quach
mai.quach3@mail.dcu.ie
School of Computing, ML-Labs,
Dublin City University, Ireland

Binh T. Nguyen
ngtbinh@hcmus.edu.vn
Department of Computer Science,
VNUHCM – University of Science

Cathal Gurrin
cathal.gurrin@dcu.ie
School of Computing, ADAPT Centre,
Dublin City University, Ireland

Graham Healy
graham.healy@dcu.ie
School of Computing, ADAPT Centre,
Dublin City University, Ireland

Abstract

Language models (LMs) are trained to excel at predicting the next word in the sequence given prior context, and humans also share this predictability in reading comprehension. Neuroscience research reveals that next-word predictability influences brain response, as recorded at millisecond resolution using electroencephalography (EEG). While our evidence indicates that advanced LMs achieve accuracies closely aligned with human performance at the next-word prediction task, this raises the question: Does higher prediction accuracy necessarily mean that these models adequately capture the cognitive signals associated with human reading comprehension? Here, we generate regressors for both humans and LMs based on two information measures, including top-1 prediction and surprisal, to predict event-related potential (ERP) elicited from EEG recordings which reflect different stages of cognitive processing during reading. We argue that modelling ERP patterns offers fine-grained analysis of the cognitive plausibility of various LMs during reading. Our results indicate that only surprisal potentially correlates with language-processing ERPs, especially for open-class words with high semantic content. Moreover, our findings challenge the assumption that scaling LMs with increased parameters and computational budgets will consistently lead to improved convergence with human-like linguistic processing.

Keywords: Cognitive Neuroscience; Reading Comprehension; Language Models; GPT; Electroencephalography (EEG)

Introduction

Word predictability reflects a broader cognitive process in which individuals continuously integrate context to anticipate upcoming events and test those predictions against perceptual input from the utterances they hear or read (Bar, 2007). During reading, predictability effects are hypothesized to reflect the cognitive demands associated with probabilistic inference, whereby the brain incrementally evaluates and updates possible upcoming word interpretations (Shain et al., 2024). From an information-theoretic perspective (Shannon, 1948), prediction serves as a core function of a probabilistic, generative cognitive system that incrementally processes the words of an unfolding sentence. Linguistic units convey quantifiable information, with measures such as surprisal (the unexpectedness of a word given its prior context). Thus, surprisal serves as a useful metric for quantifying word-by-word predictability during sentence processing (Hale, 2016).

Research has indicated that surprisal is a reliable predictor of neural responses during reading, particularly in relation to the N400 component. Michaelov & Bergen (2020) found that surprisal effectively predicts variations in N400 amplitude, a neural indicator of processing difficulty during language comprehension. Frank et al. (2013) further supported these findings by analysing EEG data from participants reading identical sentences and examining four distinct ERP components. Their results highlighted that surprisal estimates significantly predict N400 amplitude, with more surprising words eliciting larger negative N400 responses.

Surprisal modelling from LMs has been widely used to explain neural responses during language comprehension, with early work showing that trigram-based surprisal correlates with the N400 in reading studies (Frank et al., 2015; Willems et al., 2016; Armeni et al., 2019). More recent work with transformer-based LMs reports similar effects: Heilbron et al. (2019) showed that GPT-2's word-by-word unpredictability aligns with N400-like responses during audiobook listening, while Michaelov et al. (2024) found that GPT-3 surprisal best predicted N400 amplitude across models, suggesting that effects of expectancy, plausibility, and contextual semantic similarity can be explained by variations in word predictability. Although GPT-2 outperforms TransformerXL and XLNet in psycholinguistic prediction (Hao et al., 2020), evidence that larger models yield stronger predictability–processing relationships is mixed. Within the GPT family, Shain et al. (2024) challenged claims that more advanced LMs should exhibit stronger logarithmic relationships between contextual predictability and processing difficulty. They found that surprisal estimates from GPT-3 were not more "super-logarithmic" than those from smaller models like GPT-2, despite GPT-3's greater size and computational power.

Despite major advances in Natural Language Processing (NLP), LMs still lack an interpretable, mechanistic account aligned with how the human brain processes language. This raises debate about whether LMs capture aspects of human intelligence or simply produce outputs that mimic human thought (Mitchell & Krakauer, 2023). Next-word predictability is a fundamental aspect of human language processing, which importantly supports the cognitive plausibility of LMs (Keller, 2010). When it comes to thought, we should examine brain activity. During language comprehension, the hu-

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

man brain exhibits systematic patterns of neural activity that reflect ongoing predictive processing (Fitz & Chang, 2019). Therefore, rather than examining next-word prediction performance across various LMs, we investigate the relationship between next-word predictability and neural responses in natural reading contexts, especially in longer narratives.

To investigate this, we run multiple experiments. First, we calculate top-1 prediction and lexical surprisal at the word level for content and function words across three predictors: human subjects, n-gram models, and GPT-family models, using the DERCO dataset, a language resource combining EEG and next-word prediction data (Quach et al., 2024). Next, we encode neural responses using regression-based deconvolution to estimate predictability effects on neural activity. We then compare the correlations between neural response predictions derived from top-1 prediction and surprisal estimates of language models and those obtained from human cloze responses. The purpose of this comparison is to identify which model most closely mirrors human-like predictability in reading behaviour. To provide deeper insights, these correlations will be visualised within significant time windows and across significant electrode clusters.

Methodology

Stimuli and EEG Data Preparation

We utilised the DERCo dataset (Quach et al., 2024), which contains EEG recordings from 22 native English speakers while they were reading The Grimm Brothers' Fairy Tales. Two participants ("QPF42" and "USQ95") were excluded due to excessive eye movements. Additionally, word-by-word cloze probabilities were collected through a cloze procedure on Mechanical Turk crowdsourcing platform.

EEG data were recorded using a 32-channel electrode scalp following the international 10–20 system (Klem, 1999). Since the analysis used preprocessed data, the number of word-level EEG trials in the DERCo dataset's transcript was reduced due to artifact removal. All remaining words, after preprocessing, served as stimuli for encoding brain signals. The Python library IPA was used to extract parts of speech, which were then grouped into content and function words.

Information-theoretic measures

To investigate next-word predictability, we used two measures: top-1 prediction and surprisal. These measures serve as proxies for human and computational models' expectations and processing effort in reading comprehension, capturing different aspects of cognitive load associated with prediction.

Top-1 Prediction Estimation Most LMs aim to estimate a probability distribution over the vocabulary V, for the likely next-word $w_i \in V$ at position i, given the context $w_1, w_2 \ldots w_{i-1}$ containing the sequence of preceding words in text. The highest probability, also known as top-1 prediction, for the next token, is defined as follows:

$$P_{w_i} = \max_{w_i \in V} P(w_i \mid w_1, w_2, \ldots, w_{i-1}) \quad (1)$$

Surprisal Estimation Surprisal is a measure of the unexpectedness of a target word. Hale (2001) and Levy (2008) argued that the less expected a word is in a given context, the higher its surprisal. For example, "Peter won the championship. Afterward, he was in seventh ...". If readers recognise the idiom, they can guess that the missing word is "heaven." Since the word is highly predictable, it has low surprisal and conveys minimal new information.

After the first t words of the sentence, $w_{1\ldots t}$, will be processed, the identity of the upcoming word, w_{t+1}, is still unknown and can therefore be viewed as a random variable. The surprisal is defined as the negative log probability of the actual next word, given its preceding context:

$$Surprisal\ S_{w_{t+1}} = -logP(w_{t+1}|w_{1\ldots t}) \quad (2)$$

Predictors

Human Prediction Top-1 prediction refers to the highest percentage of participants who guessed the same next word. A top-1 prediction value of 100% indicates that all participants guessed the same next word, whereas 0% indicates that no participant predicted that upcoming word. This can be simply defined as the maximum cloze probability among the possible words that could appear in the upcoming position.

Lexical surprisal, by contrast, is the cloze probability of the correct next word in the transcript. It is important to note that the cloze value of the correct next word does not necessarily equal the top-1 prediction value. These values are equal only if the word is exceptionally easy to predict, meaning that the word predicted by all participants is also the correct word in the transcript.

A major issue with cloze procedures is that zero-probability responses yield undefined surprisal values. This occurs when no participant predicts the accurate target word w_i given the preceding context $w_1, w_2, \ldots, w_{i-1}$. With realistic sample sizes, words with $P(w_i|w_1, w_2, \ldots, w_{i-1}) < 0.001$ are absent from responses, motivating an expanded probability distribution to include more words. Accordingly, following Lowder et al. (2018), we replace zero cloze probabilities with half the smallest nonzero value before computing surprisal.

N-gram models In this study, we trained bigram to quadgram models using the NLTK Python package [1]. Unlike advanced language models such as transformer-based LMs (Amaratunga, 2023; Desai et al., 2023), n-gram models have a limitation in capturing very long-range dependencies. To mitigate this, we trained our n-gram models on the Fairy Tale Corpus (Lobo & de Matos, 2010), a domain-aligned dataset to DERCo, comprising 453 fairy tales from Project Gutenberg (Klein & Manning, 2002).

To mitigate overfitting, we excluded the five Grimm Brothers' Fairy Tales used in the DERCo dataset's transcript. All punctuation was removed, and the letters were converted to lowercase. To address data sparsity, we trained separate n-gram models with no smoothing, Laplace smoothing, and

[1] https://www.nltk.org/

Kneser-Ney smoothing. Each model then used a fixed-sized context window of *n - 1* words, locating all matching windows within the problem instance and counting the number of occurrences of each possible next token to calculate word probabilities for top-1 prediction and surprisal estimations.

GPT-2 and GPT-Neo Families Generative Pre-trained Transformer (GPT) (Radford, 2018) is a transformer-based autoregressive language model that uses a multi-head "attention" mechanism based on an encoder-decoder architecture (Vaswani, 2017). We selected GPT-2 and GPT-Neo families for investigation because they are mainly trained to predict the upcoming tokens based on probability in a left-to-right manner, similar to the cloze procedure conducted in next-word prediction tasks (Taylor, 1953).

The transformer models take token sequences (e.g., words, phonemes, or punctuation) as input, with context window depends on the story length. If a story exceeds one context window (e.g., 1024 tokens), probability estimates for subsequent tokens are conditioned on the second half of the previous window. Predictions were performed separately for each story, with context reset between stories rather than carrying over the context from the previous one. The story's topic was included in the prediction, as it was disclosed to participants at the beginning of the experiment in the DERCo dataset.

For words spanning multiple tokens, the word probability was calculated as the joint probability of the tokens using the chain rule. Surprisal was obtained by summing token-level surprisal values, reflecting the online experiment in which participants viewed words together with associated punctuation. In contrast, top-1 prediction focused only on correct word identification; since participants typed only the word without punctuation, the corresponding probability was averaged across the constituent tokens.

Models were implemented in PyTorch using the transformer modules from the HuggingFace Hub [2]. The examined GPT family variants differed primarily in their size, with the specific hyperparameters outlined in Table 1.

Model Name	n_{layers}	n_{head}	d_{model}	n_{params}
GPT-2 Small	12	12	768	∼124M
GPT-2 Medium	24	16	1024	∼355M
GPT-2 Large	36	20	1280	∼774M
GPT-Neo 125M	12	12	768	∼125M
GPT-Neo 1.3B	24	16	2048	∼1.3B
GPT-Neo 2.7B	32	16	2560	∼2.7B

Table 1: Hyper-parameters of GPT-2, GPT-Neo families. n_{layers}, n_{head}, d_{model}, and n_{params} respectively refer to the number of layers, number of attention heads per layer, embedding size, and number of parameters.

Brain Encoding Models

Brain encoding models (Heilbron et al., 2019; Goldstein et al., 2022) entail fitting a regressor to predict neural responses at each time point based on measures such as lexical surprisal and top-1 prediction for each participant, as used in this study. However, running separate mixed effects models for each time point, as in prior studies (Frank et al., 2013; Hao et al., 2020; Oh et al., 2024), would require estimating an excessive number of parameters, potentially leading to model overfitting through the capture of noise rather than meaningful patterns. To reduce this, we use ridge regression, introduced by Hoerl & Kennard (1970), which regularizes the model to control its variability and improve prediction reliability (Bishop & Nasrabadi, 2006).

Figure 1 illustrates the procedure for predicting EEG data using information-theoretic measures from the cloze experiment and LMs. In brief, words from the DERCo transcript were aligned with the EEG recordings from the EEG-based reading experiment (serving as the ground truth), with each word onset marked as time point 0 ms to standardise timing across trials. Two measures, top-1 prediction and surprisal, were used separately as independent variables to build regressors for predicting the EEG signal in a word-by-word, time-resolved manner. After running regressions for each measure, we estimated a series of predicted EEG amplitudes from t_{min} (-100 ms) to t_{max} (500 ms) for each subject per electrode.

We quantified the similarity between the predicted and actual EEG responses using Pearson's correlation coefficient (r) at the word level to validate the model's ability to capture neural representations. Using 5-fold cross-validation, each regressor was trained on 80% of the data to learn 600 coefficients for each combination of time point and sensor, and then tested on the remaining 20% of held-out words.

To ensure consistent regression parameters across subjects, we selected the regularisation hyperparameter λ by fitting a single model to pooled data from all participants across a log-spaced range $[10^{-5}, 10^{-4}, \ldots, 10^{5}]$. The optimal λ was chosen yielding the lowest generalisation error, as estimated via 5-fold cross-validation shuffled trials, using the Scikit-learn library (Pedregosa et al., 2011).

Significance Tests

In EEG research, conducting more statistical tests increases the probability of getting a false positive result due to random chance (Greenland et al., 2016). This issue is significantly amplified in this work, with 32 EEG sensors and 600-time points resulting in 19,200 t-values per subject. To address this, we implemented cluster-based permutation tests (Maris & Oostenveld, 2007), using 5,000 permutations per test to identify significant time windows for the encoding models.

A mass-univariate testing approach applies one-sample t-tests with a "hat" variance adjustment to compensate implausibly small variances ($\sigma = 10^{-3}$) (Ridgway et al., 2012). The resulting t-statistics were used to compute p-values, which were then corrected for multiple comparisons using the false discovery rate (FDR) (Genovese et al., 2002) for each time point and electrode to ensure statistically consistent effects.

[2] https://huggingface.co/

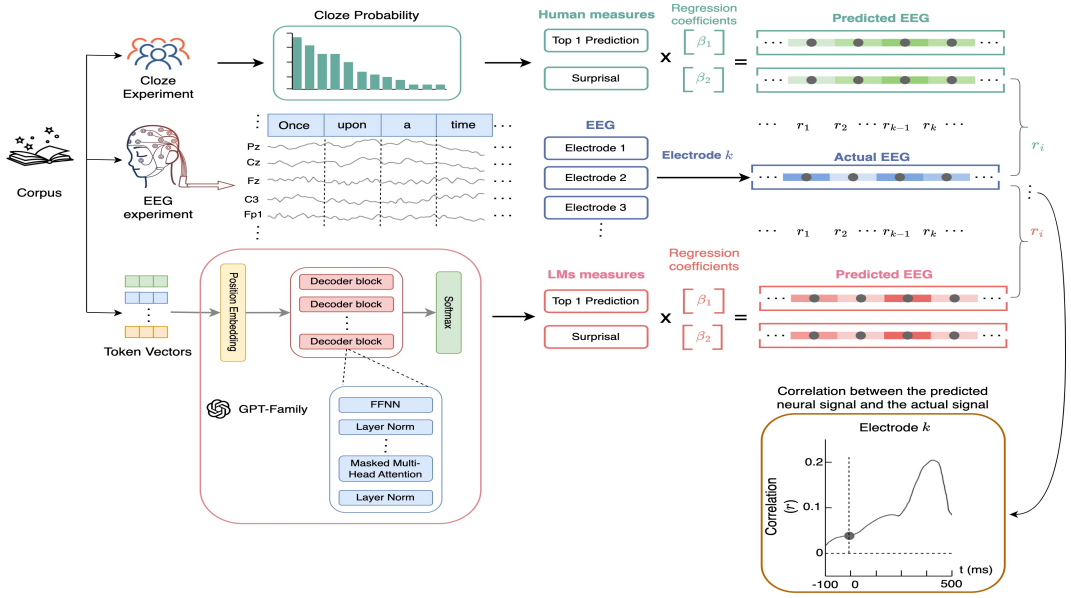

Figure 1: Brain encoding model was used to predict the neural responses to each word in the context.

Results

Performance and Human Alignment in Different Language Models

Prior studies quantified neural pattern similarity between word pairs using Pearson correlation (He et al., 2022; Goldstein et al., 2022), as it captures neural patterns similarity regardless of their amplitude. To strengthen our analysis, we also examined the Spearman correlation to provide more robust evidence for the observed relationships.

Table 2 presents the correlation between LMs and human predictions in terms of top-1 prediction and surprisal, along with their performance in the next-word prediction task, as measured by accuracy. Figure 2 further visualises differences in accuracy and joint accuracy between LMs and human predictions. These results yield several important observations.

First, humans still outperformed these LMs, achieving the highest accuracy (45.17%), highlighting the gap between LMs and human predictive capabilities. The results indicate that LMs predict next words similarly to humans in narrative contexts, with their performance gradually approaching that of humans; larger LMs are more accurate than smaller ones, but they have not surpassed human performance.

Among n-gram models, quadgrams achieved the highest accuracy, but trigrams showed the strongest correlation with humans in top-1 predictions, while bigrams had the highest surprisal correlation. Overall, correlations between n-gram models and human predictions were generally weak and inconsistent. These results show that while historically important, n-gram models struggle to capture the complexity of human-like next-word prediction.

GPT-Neo models outperformed GPT-2, with GPT-Neo 2.7B achieving the best accuracy (37.20%) and the highest percentage of joint correct predictions (29.91%) among these transformer-based LMs. Both model families demonstrated strong correlations with human behaviour in top-1 predictions and surprisal metrics, with performance improving as the model size increases. Consistent increasing in both Pearson's and Spearman's r correlations indicate a robust, distribution-independent linear relationship, suggesting larger models better capture human-like linguistic processing as evidenced by improved predictive accuracy and joint performance metrics.

Although advanced LMs improve significantly with scale, the mechanisms underlying their correlation with human behaviour remain unclear. *Do these models truly reflect human reading processes, or do they just exhibit surface-level convergence in next-word prediction?* To investigate this, we analyse and compare results from brain encoding models, using information-theoretic measures estimated by these LMs, as detailed in the following sections.

Neural Encoding Using Predictive Metrics from Human Cloze

Figure 3 shows that lexical surprisal derived from human cloze probabilities is a stronger predictor of neural responses, particularly within the N400 time window, compared to the top-1 prediction measure. These results align with the well-established semantic effects on N400 amplitude (Frank et al., 2013; Michaelov & Bergen, 2020). Additionally, encoding correlations are stronger for content words than function words, suggesting that surprisal more effectively captures neural processes associated with semantically rich lexical words (Münte et al., 2001; He et al., 2022).

Therefore, we use human cloze results as the baseline for quantifying LMs' next-word predictability, lexical surprisal as the most informative metric, and mainly focus on content words to maximise analytical sensitivity.

Model Variants	Top-1 Prediction (vs. Human)		Surprisal (vs. Human)		Top-1 Prediction (vs. Human)		Surprisal (vs. Human)		Accuracy (%)
	Pearson r	p	Pearson r	p	Spearman r	p	Spearman r	p	
Bigrams (KN)	0.09	$< p^*$	0.41	$< p^*$	0.04	0.02	0.41	$< p^*$	14.92
Trigrams (KN)	0.14	$< p^*$	0.26	$< p^*$	0.12	$< p^*$	0.28	$< p^*$	19.23
Quadgrams (KN)	0.05	0.01	0.15	$< p^*$	0.02	0.22	0.15	$< p^*$	19.36
GPT-2 Small	0.52	$< p^*$	0.73	$< p^*$	0.51	$< p^*$	0.78	$< p^*$	30.62
GPT-2 Medium	0.56	$< p^*$	0.75	$< p^*$	0.55	$< p^*$	0.80	$< p^*$	33.64
GPT-2 Large	0.59	$< p^*$	0.77	$< p^*$	0.57	$< p^*$	0.82	$< p^*$	34.59
GPT-Neo 125M	0.52	$< p^*$	0.74	$< p^*$	0.51	$< p^*$	0.78	$< p^*$	29.33
GPT-Neo 1.3B	0.59	$< p^*$	0.77	$< p^*$	0.58	$< p^*$	0.82	$< p^*$	35.77
GPT-Neo 2.7B	0.61	$< p^*$	0.77	$< p^*$	0.59	$< p^*$	0.83	$< p^*$	37.20
Human	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	45.17

Table 2: Correlation comparisons of various language model families against human benchmarks in a next-word prediction task. Metrics include accuracy, Pearson and Spearman correlation coefficients r for top-1 prediction and surprisal. Statistically significant correlations with $p^* = 0.001$ are indicated in a two-sided test.

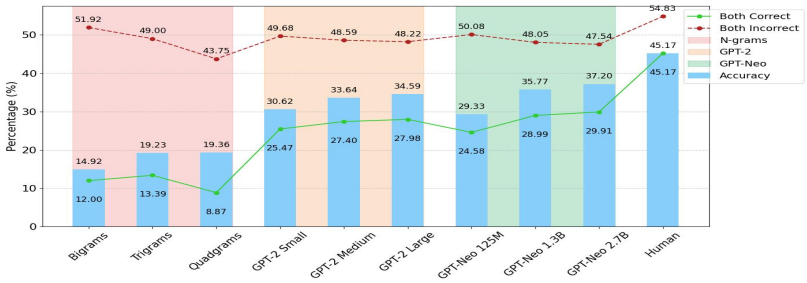

Figure 2: Performance comparison of LM families (N-grams, GPT-2, GPT-Neo) and human benchmarks in a next-word prediction task, evaluated by per-model accuracy and joint prediction percentages "both correct" (green) and "both incorrect" (red).

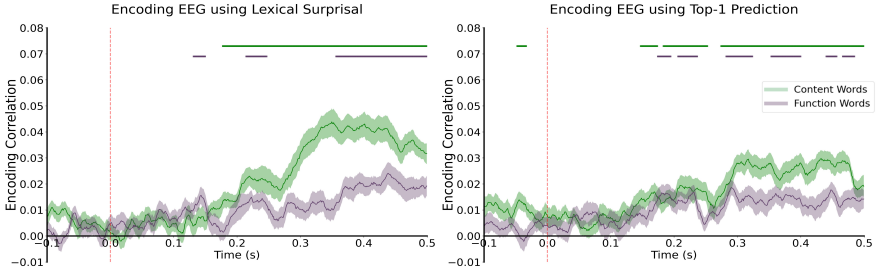

Figure 3: Correlations between human cloze–based regressors and neural responses for content and function words over time, shown for lexical surprisal (left) and top-1 prediction (right). Encoding analysis was conducted per electrode and then averaged across electrodes. Asterisks denote significant time windows ($p < 0.001$) based on cluster-based permutation tests. The shaded regions represent the between-subject standard error of the mean (SEM) of the encoding models.

Encoding Performance Comparison

We examined encoding performance at the individual-subject level and selected Bigrams, GPT-2 Large, and GPT-Neo 2.7B as representative models for each LM family based on their top performance within their groups (see Figure 2). As shown in Figure 4, the GPT-2 Large regression model shows the strongest correlations, with its mean and standard deviation for individual subjects closely aligning with human predictive models. Both the overall shape and the pattern of encoding correlation variances (i.e., the increase or decrease in values) are similar between GPT-2 Large and the human regressor.

Topographic EEG analysis

Significance levels of observed differences are reported in Figure 5. The transformer-based LMs can effectively track human neural signals using surprisal metric, particularly in predicting the content words compared to n-grams. Additionally, surprisal-based GPT-2 shows the best alignment with the encoding results from the cloze procedure.

In the early time window (*-100-100 ms*), no significant spatial pattern emerges, suggesting diffuse neural processing. However, in the *200–300 ms* window, changes in correlation are observed in the centro-frontal and parieto-occipital regions, in line with the view that the P200 component re-

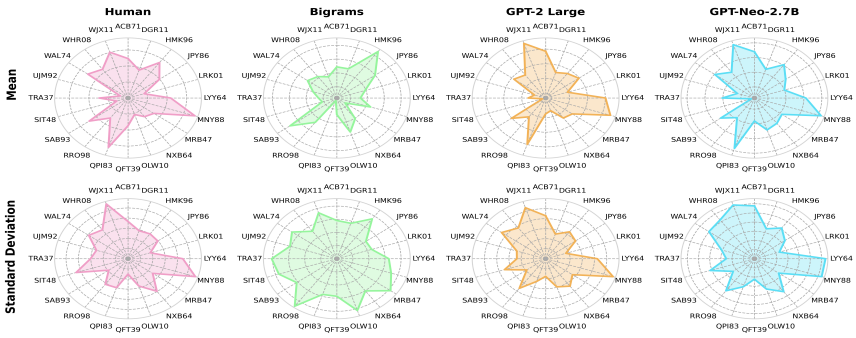

Figure 4: Radar plots of the mean (top) and standard deviation (bottom) of cross-subject surprisal correlations for content words, averaged over electrodes.

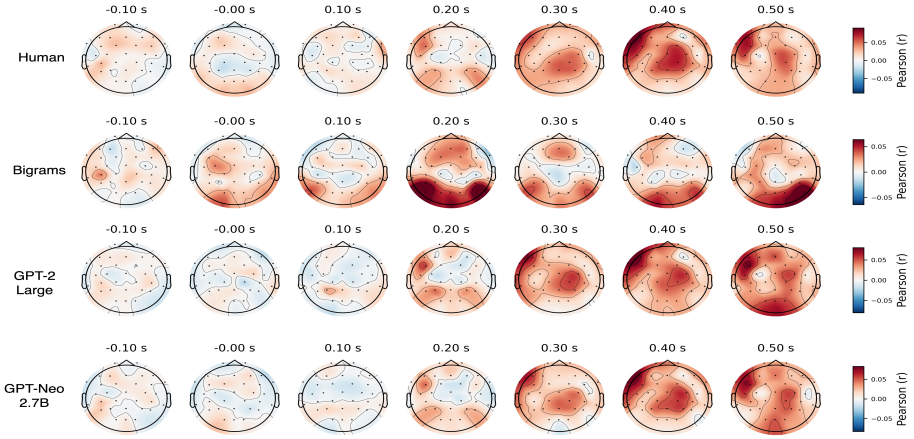

Figure 5: Full EEG Topographies of grand-averaged encoding correlations (Pearson's *r*) for lexical surprisal, computed by human predictive modelling and representative LMs over time.

flects the processing of unexpected or affectively salient linguistic information (Raney, 1993; Leuthold et al., 2015). In the *300–500* ms range, stronger encoding correlations are observed, aligning with the N400 effect, which is associated with the "expectedness" level representation on the processing of upcoming words (Hoeks et al., 2004; Ye et al., 2022).

Discussion and Conclusion

This study investigated how LM predictions relate to neural responses during reading using surprisal and top-1 prediction, and their alignment with human next-word predictability. The findings indicated that surprisal-based predictors showed significant differences in neural responses for these two lexical categories, whereas top-1 prediction failed to capture this difference. Furthermore, although larger and more advanced language models typically showed a close correspondence to human productions in next-word prediction in terms of information-theoretic measures, our results demonstrated that larger model size and increased computational resources may not reliably produce more human-like language processing. Notably, surprisal-based GPT-2 Large regression substantially outperformed larger and more advanced language models in both overall and individual subject-level analyses.

Despite these insights, several limitations must be acknowledged to guide future improvements. First, our analysis focused mainly on GPT-family transformer models; however, the rapidly evolving LM landscape includes other unidirectional architectures, such as LLaMA (Touvron et al., 2023) and DeepSeekMoE (Dai et al., 2024), whose distinct training objectives may impact cognitive modelling. Future work should therefore examine a broader range of transformer-based models. Second, we examined next-word predictability only via lexical groups, capturing a limited aspect of human reading behaviour. Prior work suggests that surprisal sensitivity also varies with word frequency and predictability (van Schijndel & Linzen, 2019; Michaelov & Bergen, 2020; Xia et al., 2023; Oh et al., 2024), potentially limiting generalisability across word difficulty levels. Finally, our analyses emphasised top-down semantic processing, whereas predictability estimates derived from language models reflect a combination of semantic and syntactic information (Qian & Levy, 2019; Wilcox et al., 2021; Arehalli et al., 2022); future work should incorporate syntactic contributions.

Acknowledgements

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6183 and 13/RC/2106_P2. We would like to thank the anonymous reviewers for their helpful remarks.

Code Availability

All the scripts for analysis can be found at https://github.com/Tayerquach/brain_encoding_model.

References

Amaratunga, T. (2023). Nlp through the ages. In *Understanding large language models: Learning their underlying concepts and technologies* (pp. 9–54). Springer.

Arehalli, S., Dillon, B. W., & Linzen, T. (2022). Syntactic surprisal from neural models predicts, but underestimates, human processing difficulty from syntactic ambiguities. In *Proceedings of the 26th conference on computational natural language learning (conll)* (pp. 301–313).

Armeni, K., Willems, R. M., Van den Bosch, A., & Schoffelen, J.-M. (2019). Frequency-specific brain dynamics related to prediction during language comprehension. *NeuroImage*, *198*, 283–295.

Bar, M. (2007). The proactive brain: using analogies and associations to generate predictions. *Trends in cognitive sciences*, *11*(7), 280–289.

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4) (No. 4). Springer.

Dai, D., Deng, C., Zhao, C., Xu, R., Gao, H., Chen, D., ... Liang, W. (2024, August). DeepSeekMoE: Towards ultimate expert specialization in mixture-of-experts language models. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1280–1297). Bangkok, Thailand: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.acl-long.70/ doi: 10.18653/v1/2024.acl-long.70

Desai, B., Patil, K., Patil, A., & Mehta, I. (2023). Large language models: A comprehensive exploration of modern ai's potential and pitfalls. *Journal of Innovative Technologies*, *6*(1).

Fitz, H., & Chang, F. (2019). Language erps reflect learning through prediction error propagation. *Cognitive Psychology*, *111*, 15–52.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2013). Word surprisal predicts n400 amplitude during reading. In *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 878–883).

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

Genovese, C. R., Lazar, N. A., & Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage*, *15*(4), 870–878.

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., ... others (2022). Shared computational principles for language processing in humans and deep language models. *Nature neuroscience*, *25*(3), 369–380.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. *European journal of epidemiology*, *31*(4), 337–350.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the north american chapter of the association for computational linguistics.*

Hale, J. (2016). Information-theoretical complexity metrics. *Language and Linguistics Compass*, *10*(9), 397–412.

Hao, Y., Mendelsohn, S., Sterneck, R., Martinez, R., & Frank, R. (2020). Probabilistic predictions of people perusing: Evaluating metrics of language model performance for psycholinguistic modeling. In *Proceedings of the workshop on cognitive modeling and computational linguistics* (pp. 75–86).

He, T., Boudewyn, M. A., Kiat, J. E., Sagae, K., & Luck, S. J. (2022). Neural correlates of word representation vectors in natural language processing models: Evidence from representational similarity analysis of event-related brain potentials. *Psychophysiology*, *59*(3), e13976.

Heilbron, M., Ehinger, B., Hagoort, P., & De Lange, F. P. (2019). Tracking naturalistic linguistic predictions with deep neural language models. In *2019 conference on cognitive computational neuroscience (ccn 2019)* (pp. 424–427).

Hoeks, J. C., Stowe, L. A., & Doedens, G. (2004). Seeing words in context: the interaction of lexical and sentence level information during reading. *Cognitive brain research*, *19*(1), 59–73.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: applications to nonorthogonal problems. *Technometrics*, *12*(1), 69–82.

Keller, F. (2010). Cognitively plausible models of human language processing. In *Proceedings of the acl 2010 conference short papers* (pp. 60–67).

Klein, D., & Manning, C. D. (2002). Fast exact inference with a factored model for natural language parsing. *Advances in neural information processing systems*, *15*.

Klem, G. H. (1999). The ten-twenty electrode system of the international federation. The international federation of clinical neurophysiology. *Electroencephalogr. Clin. Neurophysiol. Suppl.*, *52*, 3–6.

Leuthold, H., Kunkel, A., Mackenzie, I. G., & Filik, R. (2015). Online processing of moral transgressions: Erp evidence for spontaneous evaluation. *Social cognitive and affective neuroscience*, *10*(8), 1021–1029.

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

Lobo, P. V., & de Matos, D. M. (2010). Fairy tale corpus organization using latent semantic mapping and an item-to-item top-n recommendation algorithm. In *Proceedings of the seventh international conference on language resources and evaluation (lrec'10).*

Lowder, M. W., Choi, W., Ferreira, F., & Henderson, J. M. (2018). Lexical predictability during natural reading: Effects of surprisal and entropy reduction. *Cognitive science*, *42*, 1166–1183.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of eeg-and meg-data. *Journal of neuroscience methods*, *164*(1), 177–190.

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. *Neurobiology of language*, *5*(1), 107–135.

Michaelov, J. A., & Bergen, B. K. (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).

Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in ai's large language models. *Proceedings of the National Academy of Sciences*, *120*(13), e2215907120.

Münte, T. F., Wieringa, B. M., Weyerts, H., Szentkuti, A., Matzke, M., & Johannes, S. (2001). Differences in brain potentials to open and closed class words: Class and frequency effects. *Neuropsychologia*, *39*(1), 91–102.

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 *Proceedings of the 18th conference of the european chapter of the association for computational linguistics (volume 1: Long papers)* (pp. 2644–2663).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, *12*, 2825–2830.

Qian, P., & Levy, R. P. (2019). Neural language models as psycholinguistic subjects: representations of syntactic state..

Quach, B. M., Gurrin, C., & Healy, G. (2024). Derco: A dataset for human behaviour in reading comprehension using eeg. *Scientific Data*, *11*(1), 1104.

Radford, A. (2018). Improving language understanding by generative pre-training.

Raney, G. E. (1993). Monitoring changes in cognitive load during reading: an event-related brain potential and reaction time analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(1), 51.

Ridgway, G. R., Litvak, V., Flandin, G., Friston, K. J., & Penny, W. D. (2012). The problem of low variance voxels in statistical parametric mapping; a new hat avoids a 'haircut'. *Neuroimage*, *59*(3), 2131–2141.

Shain, C., Meister, C., Pimentel, T., Cotterell, R., & Levy, R. (2024). Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, *121*(10), e2307876121.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379–423.

Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism quarterly*, *30*(4), 415–433.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., ... others (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

van Schijndel, M., & Linzen, T. (2019). Can entropy explain successor surprisal effects in reading? In *Proceedings of the society for computation in linguistics (scil) 2019* (pp. 1–7).

Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

Wilcox, E., Vani, P., & Levy, R. (2021). A targeted assessment of incremental processing in neural language models and humans. 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)* (pp. 939–952).

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & Van den Bosch, A. (2016). Prediction during natural language comprehension. *Cerebral cortex*, *26*(6), 2506–2516.

Xia, M., Artetxe, M., Zhou, C., Lin, X. V., Pasunuru, R., Chen, D., ... Stoyanov, V. (2023). Training trajectories of language models across scales. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 13711–13738).

Ye, Z., Xie, X., Liu, Y., Wang, Z., Chen, X., Zhang, M., & Ma, S. (2022). Towards a better understanding of human reading comprehension with brain signals. In *Proceedings of the acm web conference 2022* (pp. 380–391).