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

Anonymous ACL submission

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). However, little is known about which measures of predictability successfully express the similarity between LMs and humans in the reading comprehension process. 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. Our results indicate that while the more advanced LMs show a close correspondence to human performance in word prediction accuracy, 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.

1 Introduction

Despite significant advances in Natural Language Processing (NLP), LMs still struggle to illustrate an adequate neurally-mechanistic picture of human language processing. This initiated a debate on whether LMs capture human intelligence or are simply called "thinking" in any human-like sense (Mitchell and Krakauer, 2023). Next-word predictability is a fundamental aspect of human language processing, which importantly supports LMs to be cognitively plausible (Keller, 2010). When it comes to thought, we need to examine brain activity. This is because when people engage in language comprehension, their brains display particular patterns of electrical activity (Fitz and Chang, 2019). Therefore, rather than examining next-word prediction performance across various LMs, we should 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 (GPT-2 and GPT-Neo), 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.

2 Background

2.1 Linguistic Prediction at the Computational Level

Word predictability effects fit into a broader picture of human cognition, in which individuals continuously integrate new input with context to make predictions about upcoming events and test those predictions against their perceptual input from the utterances they hear or read (Bar, 2007). But what cognitive processes underlie these predictability effects? One view is that predictability effects reflect the cognitive costs associated with probabilistic inference over sentence interpretations (Shain et al., 2024). This perspective, grounded in information

theory (Shannon, 1948), frames prediction as an intrinsic function of a generative, probabilistic mental processor. Under this framework, linguistic units convey quantifiable information, with measures such as surprisal (the unexpectedness of a word given its prior context). In general, surprisal serves as a useful metric for quantifying word-by-word predictability during incremental 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 and Bergen (Michaelov and Bergen, 2020) found that surprisal effectively predicts variations in N400 amplitude, a neural indicator of processing difficulty during language comprehension. Frank et al. (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. Lindborg et al. (Lindborg et al., 2023) provided additional evidence, indicating that semantic surprisal effects are specifically confined to the N400 time window (300-500 ms post-stimulus), and its effect topography closely aligns with conventional ERP analyses of expected versus unexpected words.

2.2 Neural Responses Prediction under Cloze Estimates

Unlike language models, determining the exact probability of the next word generated in the human mind remains unattainable due to the complexity and opacity of neural computations. Nevertheless, predictability in psycholinguistics is commonly studied using the cloze procedure (Taylor, 1953), a traditional approach that involves asking participants to predict and complete unfinished sentences or passages based on the accumulated preceding context. This approach is widely regarded as the gold standard for estimating human lexical probabilities, with cloze probability emerging as the primary metric for contextual word predictability (Kutas and Hillyard, 1984; Van Petten and Luka, 2012; Brothers and Kuperberg, 2021).

The cloze procedure offers several advantages. First, it indirectly reflects human subjective probabilities, capturing how individuals perceive the likelihood of specific linguistic outcomes. Second, cloze-based estimates outperform corpus-derived probabilities in predicting human reading patterns (Smith and Levy, 2011). Empirical studies consistently demonstrate that words with higher cloze probabilities elicit smaller N400 responses than words with lower cloze probabilities (Kutas and Hillyard, 1984; Kutas and Federmeier, 2011; Kuperberg et al., 2020; Brothers and Kuperberg, 2021). Furthermore, research has identified a strong linear correlation between cloze probability and lexical processing difficulty (Smith and Levy, 2013; Brothers and Kuperberg, 2021).

2.3 Neural Responses Prediction under Language Model's Probability

Surprisal modelling from LMs has been commonly applied to predict neural responses during language comprehension. Surprisal, estimated using simple and efficient trigram models, has been shown to correlate positively with the N400 effect observed in reading studies (Frank et al., 2015; Willems et al., 2016; Armeni et al., 2019). For more advanced LMs, Heilbron et al. (Heilbron et al., 2019) found that GPT-2's word-by-word un(predictability) estimates align with ERPs when participants listened to audiobooks, revealing strong negative responses at 400 ms. As further evidence, Michaelov et al. (Michaelov et al., 2024) compared transformer- and non-transformer-based LMs and found that GPT-3 surprisal best predicted N400 amplitude, suggesting that effects of expectancy, plausibility, and contextual semantic similarity can all be explained by variations in word predictability.

Within transformer-based LMs, Hao et al. (Hao et al., 2020) showed GPT-2 as the most effective in predicting psycholinguistic patterns, outperforming both TransformerXL and XLNet in terms of psycholinguistic predictive power. Their results highlighted that GPT-2 better captures lexical surprisal aligned with human predictability, potentially making it a strong candidate for psycholinguistic modelling in comparison to more complex architectures. Within the GPT family, Shain et al. (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.

3 Methodology

3.1 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.

High-density EEG data were recorded using a 32-channel electrode scalp following the international 10–20 system (Klem, 1999). Since the analysis used the 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 the parts of speech, which were then grouped into content and function words.

3.2 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 word prediction.

3.2.1 Top-1 Prediction Estimation

The objective of most LMs is to compute a probability distribution over the model's vocabulary, V, for the likely next-word $w_i \in V$ at position i, given the context $w_1, w_2 \dots w_{i-1}$ containing the sequence of preceding words in text. The highest probability, also known as top-1 prediction, for the next token, is calculated as the following formula:

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

3.2.2 Surprisal Estimation

Surprisal is a measure of the unexpectedness of a target word. Hale (Hale, 2001) and Levy (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, $w_{1\dots t}$, 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:

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

3.3 Predictors

3.3.1 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, while a value of 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 predictions result in undefined surprisal values. There are cases where the target word w_i is not predicted by any participant as a possible continuation of $w_1, w_2, \dots, w_{i-1}$. With realistic sample sizes, words with $P(w_i|w_1, w_2, \dots, w_{i-1}) < 0.001$ will be completely absent from the participants' responses. This highlights the need to expand the probability distribution to include more words. To address this, we followed the approach used by Lowder et al. (Lowder et al., 2018), which involved replacing these cloze values of zero with half the value of the lowest nonzero cloze value before converting them to surprisal values.

3.3.2 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 and de Matos, 2010), a domain-aligned

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

dataset to DERCo, comprising 453 fairy tales from Project Gutenberg (Klein and Manning, 2002).

To avoid potential overfitting, we removed the five Grimm Brothers' Fairy Tales included in the DERCo dataset's transcript. All punctuation was removed, and the letters were converted to lowercase. To address data sparsity, we trained separate models using no smoothing, Laplace smoothing, and Kneser-Ney smoothing (see Appendix A for details). Each n-gram 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 top1-prediction and surprisal estimations.

3.3.3 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) (see Appendix C.1). We selected GPT-2 and GPT-Neo families for investigation because they are 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 (see Section 2.2).

The transformer model's input is a sequence of tokens — words, phonemes, or punctuation. The number of tokens per context window depends on the story length. If a story exceeds one context window (e.g., 1024 tokens), probability estimates for the remaining tokens are conditioned on the second half of the previous window. Predictions were performed separately for each story in the transcript, with the entire preceding context restarted rather than carrying over the context from the previous story. Additionally, the story's topic was included in the prediction, as it was disclosed to participants at the beginning of the Mechanical Turk 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. Regarding surprisal estimation, we summed the surprisal values of the joint tokens, as this approach aligned with the online experiment presentation to participants, which included both the word and related punctuation. In contrast, for top-1 prediction estimation, since the focus was only on correct word prediction, participants were instructed to type only the word (excluding punctuation). Therefore, the probability used for calculating top-1 prediction was averaged over the joint tokens.

The 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 Appendix C.2.

3.4 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, 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 and Kennard (Hoerl and Kennard, 1970), which regularizes the model to control its variability and improve prediction reliability (Bishop and Nasrabadi, 2006). For more details about the ridge regression using for brain activity, please refer to Appendix E.

Figure 1 illustrates the procedure for predicting EEG data using information-theoretic measures computed by the cloze experiment and LMs. In brief, the words from the DERCo dataset's 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. The two measures, top-1 prediction and surprisal, are 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.

To quantify the similarity between the predicted and actual EEG responses, we compute the Pearson correlation coefficient (r) for each word. This correlation-based evaluation produces a set of correlation values that reflect the alignment between model-predicted and observed EEG signals, validating the model's effectiveness in capturing neural representations during reading. Using 5-fold cross-validation, we trained each regressor on 80% of the

[2] https://huggingface.co/

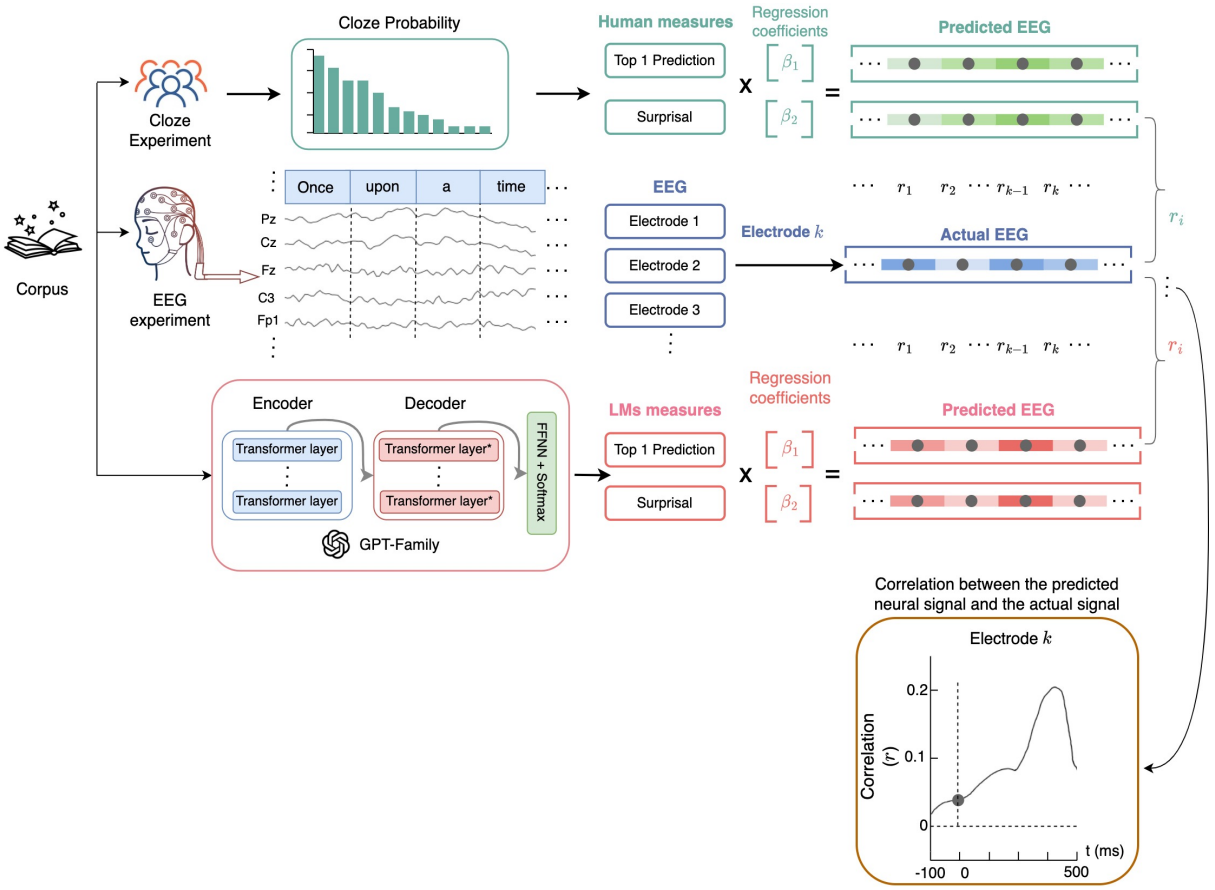

Figure 1: Brain encoding model was used to predict the neural responses to each word in the context.

dataset to learn 600 coefficients for each combination of time point and sensor, then predicted EEG activity for held-out words in the remaining 20%.

To ensure consistent regression parameters across subjects, we determined the optimal hyperparameter λ by fitting a single regression model to the data from all participants for each value of λ over a log-spaced interval $[10^{-5}, 10^{-4}, \ldots, 10^{5}]$. We then selected the λ value yielding the lowest generalisation error, as estimated via 5-fold cross-validation (with shuffling enabled) over trials, using the Scikit-learn library (Pedregosa et al., 2011).

3.5 Significance Tests

All statistical analyses used two-tailed tests with significance levels of $\alpha = 0.05, 0.01$, or 0.001, depending on the desired confidence interval. Before conducting univariate tests, we ensured that our data met two main assumptions: (1) normality and (2) absence of outliers. Normality was assessed using the D'Agostino-Pearson test, while outlier detection was based on the 1.5 IQR (interquartile range) method. If both assumptions were met, a one-sample t-test was applied as the parametric test; otherwise, non-parametric alternatives (e.g., Wilcoxon signed-rank tests) were used. This approach helps reduce the risk of misleading conclusions and statistical errors.

In EEG research, conducting more statistical tests increases the probability of getting a false positive result due to random chance (Greenland et al., 2016). In this study, the multiple comparisons problem is significantly more pronounced, 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 and 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 adjusted for multiple comparisons. Multiple comparisons were corrected using the false discovery rate (FDR) (Genovese et al., 2002) for each time point and electrode to ensure statistical

consistency of the effects.

4 Results

4.1 Performance and Human Alignment in Different Language Models

Previous 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. Please refer to Appendix B for details about these correlations.

Table 1 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 6 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%), which underscores the existing gap between LMs and human predictive capabilities. The results indicate that LMs predict the 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 correlation in surprisal. 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. The consistent increasing patterns observed in both Pearson r and Spearman r correlations indicate a robust linear relationship independent of distributional assumptions, suggesting larger models better capture human-like linguistic processing as evidenced by improved predictive accuracy and joint performance metrics (see Figure 6).

While 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.

4.2 Neural Encoding Using Predictive Metrics from Human Cloze

Figure 2 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 and Bergen, 2020; Lindborg et al., 2023). 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.

4.3 Encoding Performance Comparison

In Figure 3, transformer-based LMs significantly outperform traditional statistical models (i.e., n-grams) across all electrodes and subjects. GPT-2 Large achieves the highest performance, surpassing all other LMs. Importantly, surprisal-based regression estimates from the larger GPT-Neo 2.7B model do not provide stronger prediction correlations than those from the smaller, less accurate models (e.g., GPT-2 variants), consistent with prior research (Shain et al., 2024).

To further examine encoding performance by individual subject, we selected Bigrams, GPT-2 Large, and GPT-Neo 2.7B as representatives for each LM family, based on their top performance within their respective groups (see Section 4.1). In Figure 4, the GPT-2 Large regression model shows the strongest correlations, with its mean and standard deviation values for individual subjects closely

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 1: 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.

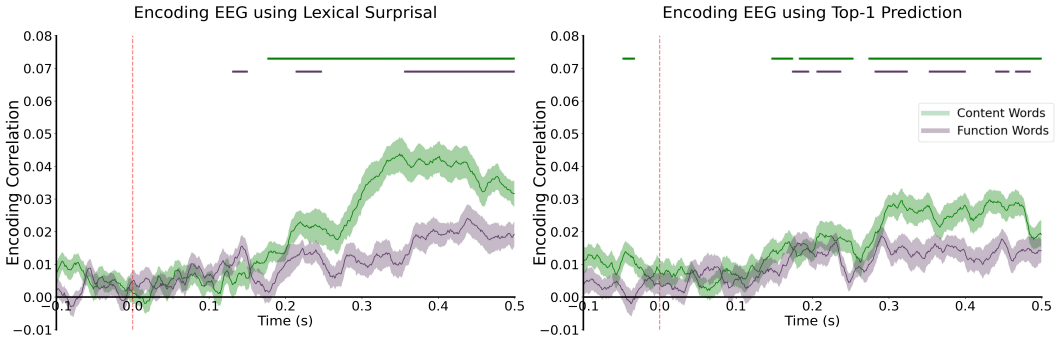

Figure 2: Correlations between regressor results from human cloze probabilities and neural responses for content and function words over time, shown for lexical surprisal (left) and top-1 prediction (right). Encoding analysis was conducted for each electrode and then averaged across electrodes. Asterisks indicate time windows at which the value is significantly different ($p < 0.001$) based on cluster-based permutation tests. The shaded regions represent the between-subject standard error of the mean (SEM) of the encoding models.

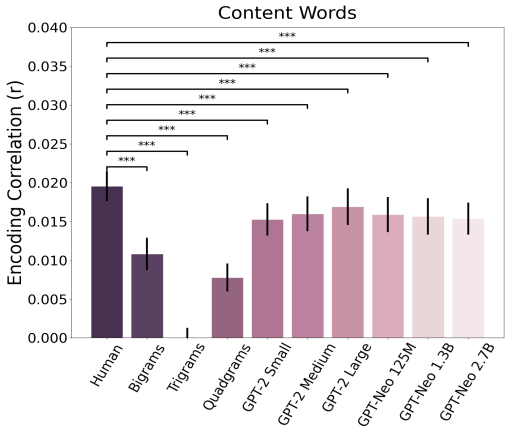

Figure 3: EEG encoding model comparison results using surprisal measure, averaged across all channels over time (from -100 to 500 ms) for all participants. Significance levels indicate statistical differences between models, with $p < 0.05$ (*), < 0.01 (**), and < 0.001 (***) based on two-tailed paired t-tests or Wilcoxon signed-rank tests. Error bars represent the SEM.

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. Detailed results for LMs and subject-wise performance are provided in Appendices F and G, respectively.

4.4 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

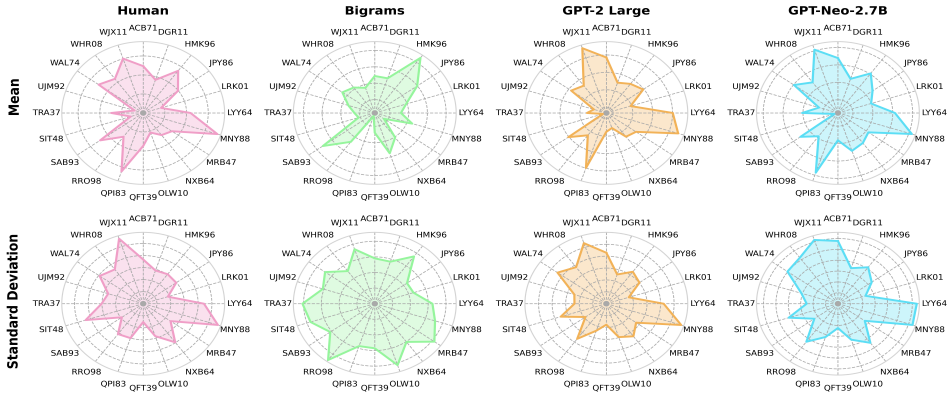

Figure 4: Radar plots of the mean (top) and standard deviation (bottom) of cross-subject surprisal correlations for content words, averaged over electrodes.

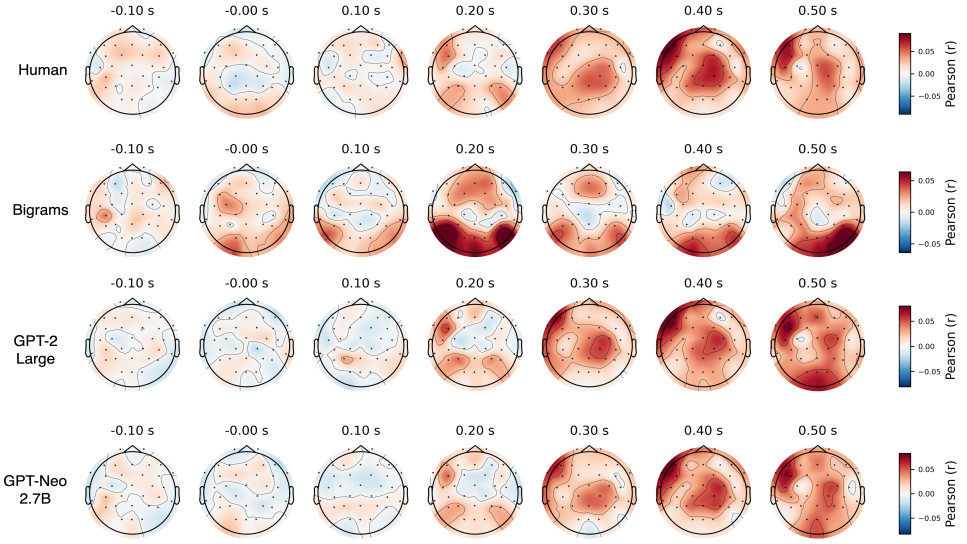

Figure 5: Full EEG Topographies of grand-averaged encoding correlations (Pearson r) for lexical surprisal, computed by human predictive modelling and representative LMs over time.

window, changes in correlation are observed in the centro-frontal and parieto-occipital regions, in line with the view that the P200 component reflects 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). For other LMs' topography, please refer to Appendix H.

5 Conclusion

This research investigated the correlation between LM predictions and neural responses during reading comprehension estimated by surprisal and top-1 prediction metrics 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, a distinction hardly captured by top-1 prediction. Furthermore, although the larger and more advanced language models typically showed a close correspondence to human productions in next-word prediction in term of information-theoretic measures, our results demonstrated that larger model size and increased computational resources may not reliably produce more human-like language processing. As shown in our experiments, surprisal-based GPT-2 Large regression substantially outperformed larger and more advanced language models in both overall and individual subject-level analyses.

Limitations

Despite the valuable insights gained, several limitations must be acknowledged to guide future improvements and broader applicability. First, our analysis primarily focused on the GPT family of transformer-based language models. While GPT models are widely popular, the landscape of transformer-based LMs has rapidly evolved. Other open-source unidirectional LLMs, such as LLaMA (Touvron et al., 2023) and DeepSeekMoE (Dai et al., 2024), also leverage transformer architectures but introduce unique training objectives and capabilities. These differences may significantly impact cognitive modelling. To develop more comprehensive insights into language model generation, future studies should broaden the scope to include a diverse set of transformer-based models.

Secondly, we examined next-word predictability only through the lens of lexical groups. While our results provided valuable insights, they represent a limited aspect of human reading behaviour. Evidence suggests that human sensitivity to surprisal extends not only to highly predictable words (van Schijndel and Linzen, 2019; Michaelov and Bergen, 2020) but also to frequent words (Xia et al., 2023; Oh et al., 2024). Consequently, this limited scope may affect the generalizability of the surprisal effect across varying levels of word difficulty.

Finally, we focused our experiments on top-down processes, specifically those related to semantic processing. However, estimates of word predictability costs derived from language models also includes contributions from bottom-up processes, particularly those driven by syntactic structures (Qian and Levy, 2019; Wilcox et al., 2021; Arehalli et al., 2022). As ongoing research, we are incorporating syntactic processing to develop more cognitively plausible predictions of neural responses based on language models.

References

Thimira Amaratunga. 2023. Nlp through the ages. In *Understanding Large Language Models: Learning Their Underlying Concepts and Technologies*, pages 9–54. Springer.

Suhas Arehalli, Brian W Dillon, and Tal Linzen. 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)*, pages 301–313.

Kristijan Armeni, Roel M Willems, Antal Van den Bosch, and Jan-Mathijs Schoffelen. 2019. Frequency-specific brain dynamics related to prediction during language comprehension. *NeuroImage*, 198:283–295.

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

Christopher M Bishop and Nasser M Nasrabadi. 2006. *Pattern recognition and machine learning*, volume 4. Springer.

Douglas G Bonett and Thomas A Wright. 2000. Sample size requirements for estimating pearson, kendall and spearman correlations. *Psychometrika*, 65:23–28.

JD Bonita, LCC Ambolode, BM Rosenberg, CJ Cellucci, TAA Watanabe, PE Rapp, and AM Albano. 2014. Time domain measures of inter-channel eeg correlations: a comparison of linear, nonparametric and nonlinear measures. *Cognitive neurodynamics*, 8:1–15.

Trevor Brothers and Gina R Kuperberg. 2021. Word predictability effects are linear, not logarithmic: Implications for probabilistic models of sentence comprehension. *Journal of Memory and Language*, 116:104174.

Damai Dai, Chengqi Deng, Chenggang Zhao, R.x. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y.k. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. 2024. DeepSeekMoE: Towards ultimate expert specialization in mixture-of-experts language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1280–1297, Bangkok, Thailand. Association for Computational Linguistics.

Bhavin Desai, Kapil Patil, Asit Patil, and Ishita Mehta. 2023. Large language models: A comprehensive exploration of modern ai's potential and pitfalls. *Journal of Innovative Technologies*, 6(1).

Hartmut Fitz and Franklin Chang. 2019. Language erps reflect learning through prediction error propagation. *Cognitive Psychology*, 111:15–52.

Stefan L Frank, Leun J Otten, Giulia Galli, and Gabriella Vigliocco. 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)*, pages 878–883.

Stefan L Frank, Leun J Otten, Giulia Galli, and Gabriella Vigliocco. 2015. The erp response to the amount of information conveyed by words in sentences. *Brain and language*, 140:1–11.

Thomas D Gauthier. 2001. Detecting trends using spearman's rank correlation coefficient. *Environmental forensics*, 2(4):359–362.

Christopher R Genovese, Nicole A Lazar, and Thomas Nichols. 2002. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage*, 15(4):870–878.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. 2022. Shared computational principles for language processing in humans and deep language models. *Nature neuroscience*, 25(3):369–380.

Sander Greenland, Stephen J Senn, Kenneth J Rothman, John B Carlin, Charles Poole, Steven N Goodman, and Douglas G Altman. 2016. Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. *European journal of epidemiology*, 31(4):337–350.

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

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

Yiding Hao, Simon Mendelsohn, Rachel Sterneck, Randi Martinez, and Robert Frank. 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*, pages 75–86.

Taiqi He, Megan A Boudewyn, John E Kiat, Kenji Sagae, and Steven J Luck. 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.

Micha Heilbron, Benedikt Ehinger, Peter Hagoort, and Floris P De Lange. 2019. Tracking naturalistic linguistic predictions with deep neural language models. In *2019 Conference on Cognitive Computational Neuroscience (CCN 2019)*, pages 424–427.

John CJ Hoeks, Laurie A Stowe, and Gina Doedens. 2004. Seeing words in context: the interaction of lexical and sentence level information during reading. *Cognitive brain research*, 19(1):59–73.

Arthur E Hoerl and Robert W Kennard. 1970. Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12(1):69–82.

Frank Keller. 2010. Cognitively plausible models of human language processing. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 60–67.

Dan Klein and Christopher D Manning. 2002. Fast exact inference with a factored model for natural language parsing. *Advances in neural information processing systems*, 15.

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

Gina R Kuperberg, Trevor Brothers, and Edward W Wlotko. 2020. A tale of two positivities and the n400: Distinct neural signatures are evoked by confirmed and violated predictions at different levels of representation. *Journal of cognitive neuroscience*, 32(1):12–35.

Marta Kutas and Kara D Federmeier. 2011. Thirty years and counting: finding meaning in the n400 component of the event-related brain potential (erp). *Annual review of psychology*, 62:621–647.

Marta Kutas and Steven A Hillyard. 1984. Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307(5947):161–163.

Tom Dupré La Tour, Michael Eickenberg, Anwar O Nunez-Elizalde, and Jack L Gallant. 2022. Feature-space selection with banded ridge regression. *NeuroImage*, 264:119728.

Hartmut Leuthold, Angelika Kunkel, Ian G Mackenzie, and Ruth Filik. 2015. Online processing of moral transgressions: Erp evidence for spontaneous evaluation. *Social cognitive and affective neuroscience*, 10(8):1021–1029.

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

Alma Lindborg, Lea Musiolek, Dirk Ostwald, and Milena Rabovsky. 2023. Semantic surprise predicts the n400 brain potential. *Neuroimage: Reports*, 3(1):100161.

Paula Vaz Lobo and David Martins de Matos. 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)*.

Matthew W Lowder, Wonil Choi, Fernanda Ferreira, and John M Henderson. 2018. Lexical predictability during natural reading: Effects of surprisal and entropy reduction. *Cognitive science*, 42:1166–1183.

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

James Michaelov and Benjamin Bergen. 2020. How well does surprisal explain n400 amplitude under different experimental conditions? In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 652–663.

James A Michaelov, Megan D Bardolph, Cyma K Van Petten, Benjamin K Bergen, and Seana Coulson. 2024. Strong prediction: Language model surprisal explains multiple n400 effects. *Neurobiology of language*, 5(1):107–135.

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

Thomas F Münte, Bernardina M Wieringa, Helga Weyerts, Andras Szentkuti, Mike Matzke, and Sönke Johannes. 2001. Differences in brain potentials to open and closed class words: Class and frequency effects. *Neuropsychologia*, 39(1):91–102.

Byung-Doh Oh, Shisen Yue, and William Schuler. 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)*, pages 2644–2663.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Peng Qian and Roger P Levy. 2019. Neural language models as psycholinguistic subjects: representations of syntactic state. Association for Computational Linguistics.

Boi Mai Quach, Cathal Gurrin, and Graham Healy. 2024. Derco: A dataset for human behaviour in reading comprehension using eeg. *Scientific Data*, 11(1):1104.

R Quian Quiroga, A Kraskov, T Kreuz, and Peter Grassberger. 2002. Performance of different synchronization measures in real data: a case study on electroencephalographic signals. *Physical Review E*, 65(4):041903.

Alec Radford. 2018. Improving language understanding by generative pre-training.

Gary E Raney. 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.

Gerard R Ridgway, Vladimir Litvak, Guillaume Flandin, Karl J Friston, and Will D Penny. 2012. The problem of low variance voxels in statistical parametric mapping; a new hat avoids a 'haircut'. *Neuroimage*, 59(3):2131–2141.

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. 2024. Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, 121(10):e2307876121.

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

Nathaniel Smith and Roger Levy. 2011. Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33.

Nathaniel J Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Cyma Van Petten and Barbara J Luka. 2012. Prediction during language comprehension: Benefits, costs, and erp components. *International journal of psychophysiology*, 83(2):176–190.

Marten van Schijndel and Tal Linzen. 2019. Can entropy explain successor surprisal effects in reading? In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pages 1–7.

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

Ethan Wilcox, Pranali Vani, and Roger Levy. 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)*, pages 939–952.

Roel M Willems, Stefan L Frank, Annabel D Nijhof, Peter Hagoort, and Antal Van den Bosch. 2016. Prediction during natural language comprehension. *Cerebral cortex*, 26(6):2506–2516.

Mengzhou Xia, Mikel Artetxe, Chunting Zhou, Xi Victoria Lin, Ramakanth Pasunuru, Danqi Chen, Luke Zettlemoyer, and Veselin Stoyanov. 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)*, pages 13711–13738.

Jiaqi Ye, Chengwei Xiao, Rui Máximo Esteves, and Chunming Rong. 2015. Time series similarity evaluation based on spearman's correlation coefficients and distance measures. In *Cloud Computing and Big Data: Second International Conference, CloudCom-Asia 2015, Huangshan, China, June 17-19, 2015, Revised Selected Papers 2*, pages 319–331. Springer.

Ziyi Ye, Xiaohui Xie, Yiqun Liu, Zhihong Wang, Xuesong Chen, Min Zhang, and Shaoping Ma. 2022. Towards a better understanding of human reading

comprehension with brain signals. In *Proceedings of the ACM Web Conference 2022*, pages 380–391.

Appendices

A Smoothing Techniques in N-Grams

When dealing with **n-gram** models, smoothing refers to the practice of adjusting empirical probability estimates to account for insufficient data.

A.1 Laplace Smoothing

Laplace smoothing, also known as add-one smoothing, assumes that each n-gram in a corpus occurs exactly one more time than it actually does. The simplest way to do this smoothing is to add one to all the n-gram counts, before we normalize them into probabilities. All the counts that used to be zero will now have a count of 1, the counts of 1 will be 2, and so on. In the equation below, we use the notation $w_i^j, i < j$, to denote the $(j-i)$-gram $(w_i, w_{i+1}, \ldots, w_j)$.

$$p(w_i \mid w_{i-n+1}^{i-1}) = \frac{1 + c(w_{i-n+1}^{i})}{|V| + \sum_{w_i} c(w_{i-n+1}^{i})}$$

where $c(a)$ denotes the empirical count of the n-gram a in the corpus, and $|V|$ corresponds to the number of unique n-grams in the corpus.

A.2 Kneser–Ney Smoothing

Unlike Laplace smoothing, Kneser-Ney smoothing not only accounts for the frequency of observed n-grams but also considers the diversity of contexts the word w has appeared in. For example, a word like "the" will have a high raw frequency but occurs in many repetitive contexts (e.g., "the dog," "the cat"). In contrast, a word like "Francisco" might appear less frequently but in more unique contexts i.e., "San Francisco", making it more informative.

This method is an extension of absolute discounting with a clever way of constructing the lower-order (backoff) model. The lower-order model is significant only when the count is small or zero in the higher-order model, and so it should be optimized for that purpose. The probability of word w_i given its previous context w_{i-n+1}^{i-1} (the previous $n-1$ words) using Kneser-Ney smoothing is expressed as follows:

$$P_{KN}(w_i \mid w_{i-n+1}^{i-1}) = \frac{\max\left(c_{KN}(w_{i-n+1}^{i}) - d, 0\right)}{c_{KN}(w_{i-n+1}^{i-1})} + \lambda(w_{i-n+1}^{i-1}) P_{KN}(w_i \mid w_{i-n+2}^{i-1}) \quad (3)$$

where,

$c_{KN}(w_{i-n+1}^{i-1})$: represents the count of the preceding context w_{i-n+1}^{i-1}. The definition of the count c_{KN} depends on whether we are counting the highest-order N-gram being interpolated (for example, trigram if we are interpolating trigram, bigram, and unigram) or one of the lower-order N-grams (bigram or unigram if we are interpolating trigram, bigram, and unigram):

$$c_{KN}(\cdot) = \begin{cases} \text{count}(\cdot) & \text{for the highest order,} \\ \text{continuation count}(\cdot) & \text{for lower orders.} \end{cases} \quad (4)$$

$\max\left(c_{KN}(w_{i-n+1}^{i}) - d, 0\right)$: applies a discount d of the observed count $c_{KN}(w_{i-n+1}^{i})$. If the discounted value becomes negative, the max function ensures it is clipped to zero.

$\lambda(w_{i-n+1}^{i-1})$: is known as the back-off weight. It is simply the amount of probability mass we left for the next lower-order model.

$P_{KN}(w_i \mid w_{i-n+2}^{i-1})$: represents the backoff probability, calculated recursively for a lower-order $n-1$-gram model.

B Correlation

Various methods have been developed to quantify relationships between time series, particularly in EEG data analysis (Quiroga et al., 2002; Bonita et al., 2014). The two most widely used correlation techniques are Pearson correlation, which measures linear relationships, and Spearman correlation, which captures monotonic relationships (Bonita et al., 2014).

B.1 Pearson Correlation

The Pearson r correlation is one of the most widely used statistical measures to quantify the dependence between pairs of time series, particularly in the analysis of bio-signals (Bonett and Wright, 2000). For instance, if we aim to measure the relationship between two signals, the Pearson r correlation quantifies the degree of relationship between the two. The correlation coefficient ranges from -1 to 1, where -1 indicates a perfect negative linear correlation, 0 signifies no correlation, and +1 represents a perfect positive linear correlation.

A point-biserial correlation is conducted with the Pearson correlation formula, except that one of the variables is dichotomous. The Pearson correlation coefficient, denoted as r_{xy}, is defined as:

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

where:

- r_{xy}: Pearson correlation coefficient between x and y
- n: number of observations
- x_i: value of x (for the i-th observation)
- y_i: value of y (for the i-th observation)

B.2 Spearman Correlation

Spearman rank correlation is a non-parametric test that is used to measure of the strength of a monotonic relationship between two independent variables (Gauthier, 2001). Compared to the Pearson correlation coefficient, the Spearman correlation coefficient operates on the ranks of the data rather than the raw data, and it does not require the relationship between variables to be linear. Since it is based on the ranks of the data, it can well represent the similarity of the trend of the time series.

Spearman correlation analysis ranks each variable separately from lowest to highest and records the difference between the ranks of each data pair (Ye et al., 2015). The sum of the square of the difference between ranks denotes the strength of the correlation between variables. If the data is strongly correlated, then the sum will be small, and vice versa. Besides, the magnitude of the sum is related to the significance of the correlation. The Spearman ranks correlation coefficient can be calculated using the following equations:

$$r_s = 1 - \frac{6\sum d_i^2}{N\left(N^2 - 1\right)} \tag{5}$$

where d_i is the difference between ranks for each data pair, and N is the number of data pairs.

C Transformer-Based Approach

C.1 Explanation of Transformer Architecture

In the transformer-based architecture, input tokens $U = (u_{i-k}, \ldots, u_{i-1})$ are first mapped through a token embedding matrix W_e. After that, a position embedding W_p is added corresponding to each word vector, resulting in the first hidden layer:

$$h_0 = UW_e + W_p. \tag{6}$$

Activities are then passed through a stack of transformer blocks consisting of a multi-headed self-attention layer, a position-wise feedforward layer, and layer normalisation. This is repeated n times for each block b, after which (log) probabilities are obtained from a (log) softmax over the transposed token embedding of h_n:

$$h_b = \text{transformer block}\ (h_{b-1}) \quad \forall i \in [1, n] \tag{7}$$

In Equation 7, each transformer block takes the output from the previous block h_{b-1} and produces a new hidden state h_b.

$$P(u_i|U) = \text{softmax}\left(h_n W_e^\top\right) \tag{8}$$

As in Equation 8, the final hidden state h_n is projected back to the vocabulary space by multiplying it with the transposed token embedding matrix $W_e^\top$. A softmax function is then applied to obtain the probability distribution $P(u_i|U)$ over the vocabulary for the next token u_i given the input sequence token U.

C.2 Capabilities of GPT Families

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 2: 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.

D Accuracy Performance

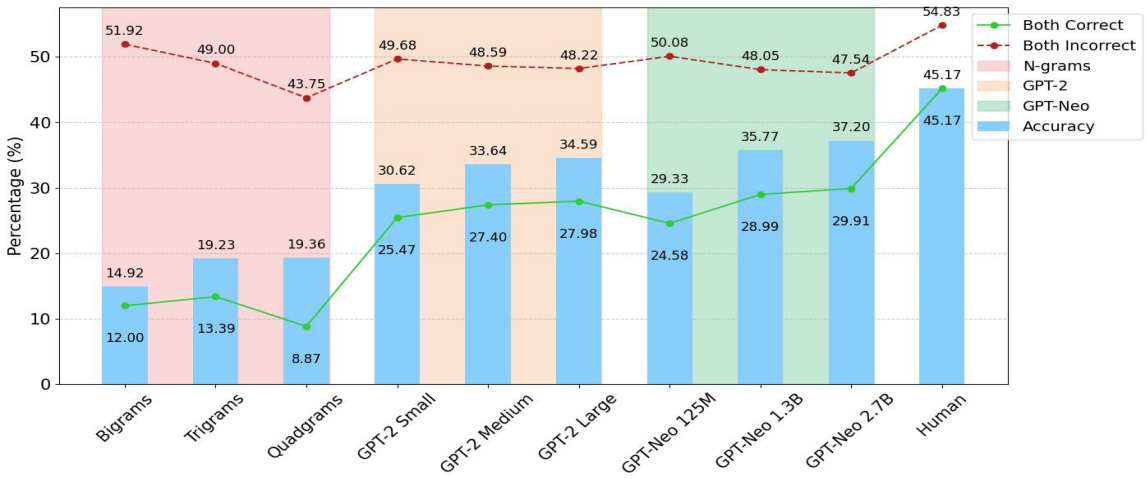

Figure 6: Performance comparison of various language model families (N-grams, GPT-2, GPT-Neo) and human benchmarks in a next-word prediction task, evaluated by per-model accuracy and joint prediction percentages "both correct" (green line) and "both incorrect" (red line).

E Brain Encoding Models using Ridge Regression

Compared to traditional least mean square regression, ridge regression provides better generalisation to unseen data by regularising the coefficient estimates, particularly in the presence of a large number of predictor variables. It is formulated as the following optimization problem, solving for the regression coefficients b^* independently at each electrode:

$$b^* = \arg\min_{b \in \mathbb{R}^p} \left(\|y - Xb\|_2^2 + \lambda \|b\|_2^2 \right), \quad (9)$$

where $X \in \mathbb{R}^{n \times p}$ represents the stimulus feature matrix, where n corresponds to the number of time samples and p to the number of features. The target vector $y \in \mathbb{R}^n$ denotes EEG data with n time points at a single electrode. The ℓ^2 norm $\| \cdot \|_2$ regularises the model's coefficients with the hyper-parameter λ controlling the penalty weight in the loss function. A low λ may lead to overfitting EEG data, while a high λ may cause underfitting of the brain encoding model (La Tour et al., 2022).

F Single-Subject Model Comparison

Figure 7: Surprisal-based encoding model comparison results for **content words** across all channels for individual participants. Significance levels indicate statistical differences between models, with $p < 0.05(*)$, $p < 0.01(**)$, and $p < 0.001(***)$ based on two-tailed paired t-tests or Wilcoxon signed-rank tests. Error bars represent the SEM.

G Cross-Subject Encoding Performance

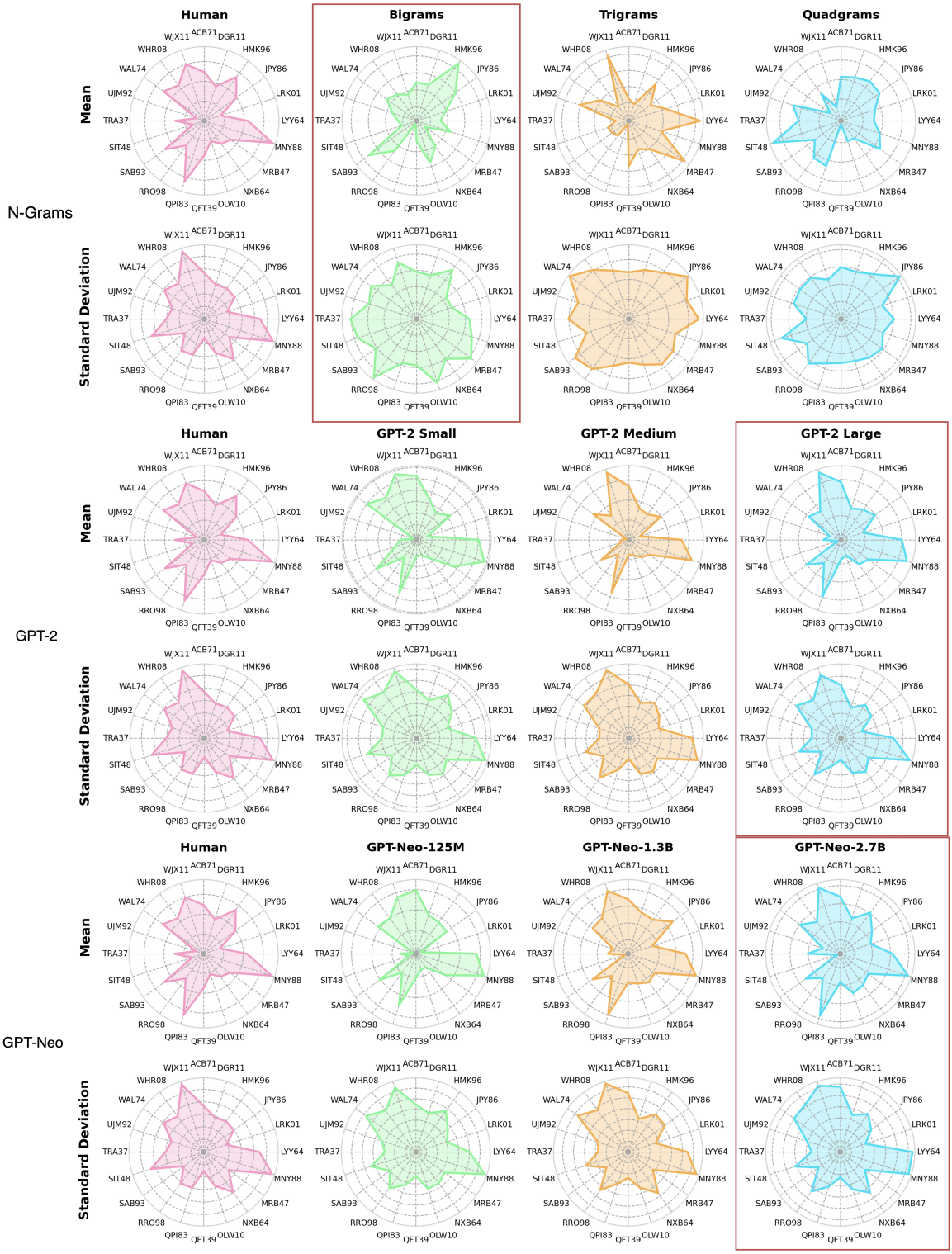

Figure 8: Radar plots of cross-subject surprisal correlations for content words across electrodes in all examined LMs.

H Cross-Model Topography

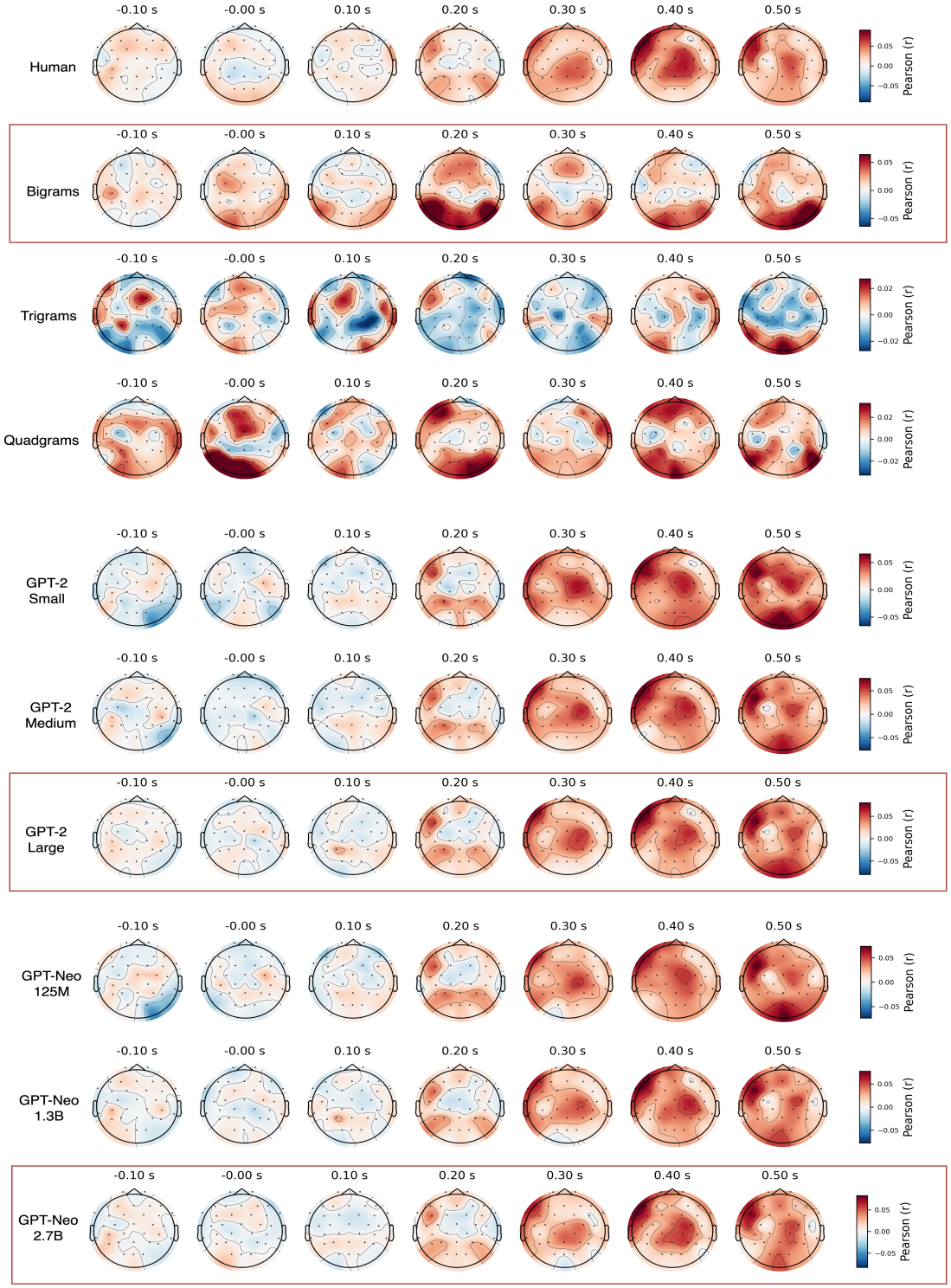

Figure 9: Full EEG Topographies of grand-averaged encoding correlations (Pearson r) for lexical surprisal, computed by human predictive modelling and all examined LMs over time.