Shared Computational Principles for Language Processing in Humans and Deep Language Models: A Partial Replication and Novel Analysis

Afifah Kashif
University of Cambridge
Centre for Human Inspired AI

Abstract

Goldstein et al. (2022) showed that autoregressive deep language models and the human brain share three computational principles during speech comprehension: next-word prediction, post-onset surprise, and contextual representation. This work replicates Figures 4-6 using the Podcast ECoG dataset (9 subjects, 1,330 electrodes) with an independently implemented pipeline using OLS encoding with 10-fold cross-validation. We reproduce all three qualitative findings, then test a novel hypothesis through variance partitioning and find that surprisal and contextual embeddings capture largely independent neural signals (8.5% shared variance), with contextual embeddings accounting for 84.7% of uniquely explained variance at peak lag.

1 Introduction

Autoregressive deep language models (DLMs) learn language through a single objective of predicting the next word given prior context. Goldstein et al. [2] used electrocorticography (ECoG) from nine participants listening to a 30-minute podcast to show that this objective portrays three principles shared with the human brain. These include continuous next-word prediction before word onset, post-onset prediction error, and contextual word representation.

This matters beyond the specific dataset. If the brain and DLMs converge on similar processing strategies despite fundamentally different architectures, next-word prediction may reflect a core strategy of biological language processing. This fits with predictive coding accounts, which frame the brain as a system that continuously generates and updates predictions to minimize surprise.

This work has two goals. First, we replicate Figures 4-6 using the publicly available Podcast ECoG dataset [4]. Second, we test a novel hypothesis that prediction error (surprisal) and contextual representations capture partially independent neural signals, quantified through variance partitioning.

2 Part 1: Replication

2.1 Methods

2.1.1 Dataset and Preprocessing

We used the Podcast ECoG dataset: nine participants, 1,330 total electrodes, preprocessed high-gamma band power (70-200 Hz) at 512 Hz. The transcript contains 5,136 words with onset times.

2.1.2 Electrode Selection

We performed a two-stage data-driven electrode selection, as we frankly we somehow missed finding the class provided significant electrode file. First, for each electrode, we compared mean high-gamma

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

power in a post-onset window (0-500 ms) against a baseline (-500-0 ms) using Wilcoxon signed-rank tests with FDR correction ($q < 0.01$) [1]. This gave 337 word-responsive electrodes across all nine subjects. We then identified 204 of these as having significant GloVe encoding through permutation testing (50 permutations, FDR $q < 0.01$). The original study identified 160 electrodes using 5,000 phase-randomized permutations. We used fewer permutations due to computational constraints, which upon further reflection could have been an entirely avoided problem. For Figure 5, we restricted analyses to the 204 GloVe-significant electrodes, using the same electrode subset across analyses.

2.1.3 Linguistic Features

For Figure 4a, we used 50-dimensional GloVe embeddings [3] (4,934 of 5,136 words matched) with two controls:

- Arbitrary Embeddings: fixed random vectors from a uniform $[-1, 1]$ distribution per word type, matching the original paper
- Shuffled GloVe: permuted word assignments

For Figure 4b, we split words into correctly predicted (top-5, rank < 5, $n = 2{,}604$) and incorrectly predicted (rank ≥ 5, $n = 2{,}532$). For incorrect words, we used GloVe embeddings of GPT-2's top-1 predicted word.

For Figure 5, we computed word-level surprisal as $-\log_2 P(w_t \mid \text{context})$, summing across sub-tokens for multi-token words, and extracted GPT-2's pre-word entropy as a measure of prediction uncertainty.

For Figure 6, we extracted contextual embeddings from GPT-2 XL's final hidden layer (layer 47 of 48), as the original specifies "the final [layer], before the classification layer." We took the last sub-token per word and reduced to 50 dimensions via PCA. Two controls were constructed:

- Averaged-Context Embeddings: mean across all occurrences of each word type, removing context-specificity
- Shuffled-Context Embeddings: permuted across occurrences of the same word type, preserving word-level statistics but destroying context-specific assignments

Following the original, we restricted this analysis to words with at least five repetitions.

2.1.4 Encoding Model

We used ordinary least-squares regression with 10-fold cross-validation, matching the original paper. Neural activity was extracted in a 200 ms window at each of 161 lags from -2,000 to +2,000 ms (25 ms steps). The encoding score is the Pearson correlation between held-out predictions and actual responses. For efficiency, we solved all electrodes simultaneously per fold via the normal equations, which is mathematically equivalent to per-electrode OLS.

For Figure 5b, we computed partial correlations between entropy and neural activity controlling for surprisal and between surprisal and neural activity controlling for entropy. This separates each predictor's unique contribution, following the original.

2.2 Results

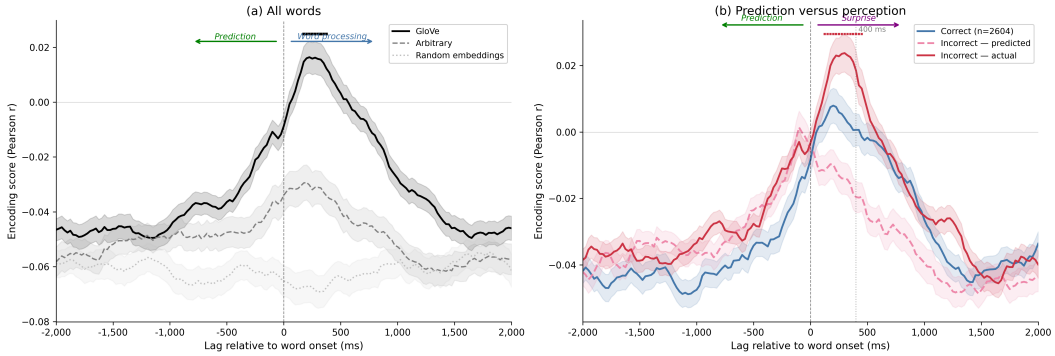

Figure 4: (a) Encoding performance for GloVe, arbitrary, and shuffled embeddings across lags. (b) Encoding for correctly predicted words, GPT-2's predicted word (incorrect trials), and the actual perceived word (incorrect trials). Top-5 accuracy split.

Figure 4a. GloVe encoding peaked at $r = 0.016$ at +225 ms, with 204 of 337 electrodes significant. Both controls remained near zero, confirming that the signal depends on semantic content rather than word identity or temporal structure. Encoding rose before word onset, consistent with anticipatory prediction.

Figure 4b. For incorrectly predicted words, encoding of GPT-2's predicted word and the actual word diverged around word onset. After onset, encoding of the perceived word clearly dominated over encoding of the predicted word. While the pre-onset separation is weaker than in the original, the post-onset pattern replicates the finding that the brain updates from prediction to perception around word onset.

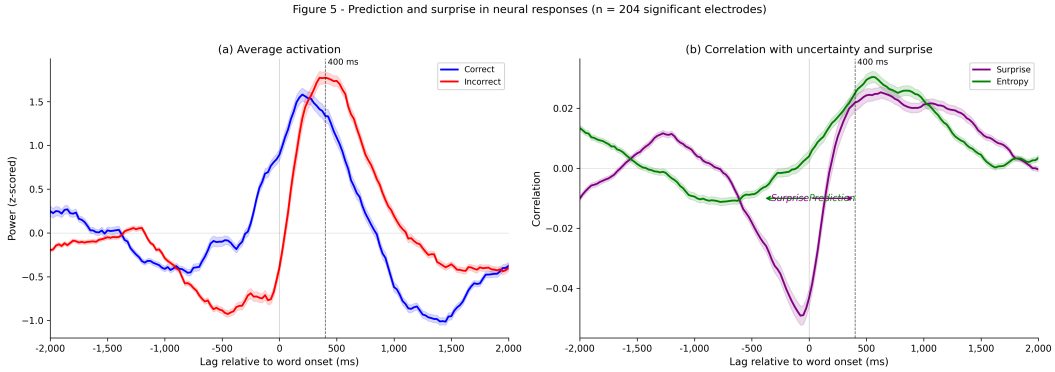

Figure 5: (a) Average high-gamma activation for correctly and incorrectly predicted words (204 GloVe-significant electrodes, z-scored). (b) Partial correlations between neural activity and entropy (controlling for surprisal, green) and surprisal (controlling for entropy, purple).

Figure 5. Panel (a) shows higher high-gamma activation for incorrectly predicted words after onset, peaking around 400 ms across the 204 GloVe-significant electrodes, consistent with the N400-like surprise response. Panel (b) shows the temporal dissociation between confidence and surprise through partial correlations: entropy correlated with pre-onset activity, while surprisal correlated with post-onset activity. This replicates the coupling between GPT-2's internal estimates and neural signals.

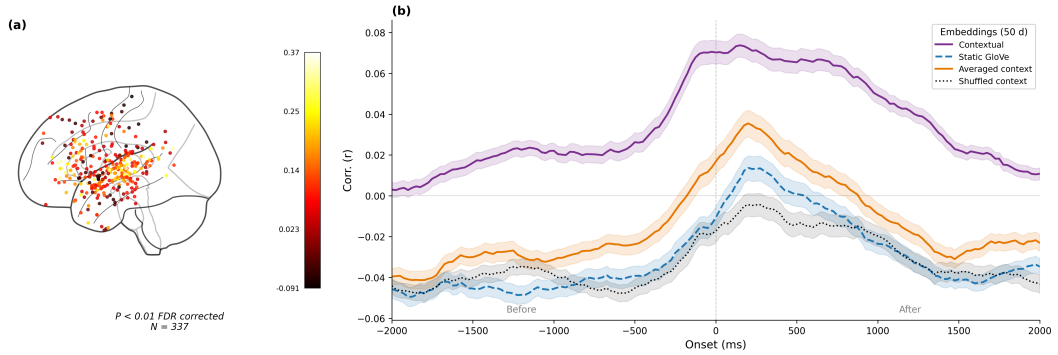

Figure 6: (a) Brain map showing peak contextual encoding strength per electrode. (b) Encoding for contextual (GPT-2 layer 47), static (GloVe), averaged-context, and shuffled-context embeddings. Restricted to words with $\geq$5 repetitions.

Figure 6. Contextual embeddings peaked at $r = 0.074$ at +150 ms, outperforming static GloVe ($r = 0.014$). Averaged-context embeddings ($r = 0.036$) reduced performance toward GloVe levels, which shows that the advantage comes from sensitivity to the specific preceding context rather than differences in training corpus. Shuffled-context embeddings performed near zero, as expected when context-specific assignments are destroyed while word-level statistics are preserved. The brain map showed strongest encoding in lateral temporal and perisylvian cortex, matching the spatial distribution reported in the original.

2.3 Discussion

All three principles were qualitatively reproduced. Encoding correlations were consistently lower than the original ($r \approx 0.01$-0.07 vs. $r \approx 0.10$-0.20). The key explanation is electrode selection. We retained 337 electrodes, many outside canonical language areas, whereas the original used 160–208 curated language-responsive electrodes concentrated in high-signal regions. Averaging over a larger set that includes non-language electrodes dilutes the mean encoding score without affecting the qualitative pattern.

Despite these quantitative differences, the core findings hold. The brain encodes predicted words before onset, registers prediction error after onset with a temporal profile that dissociates uncertainty from surprise, and represents words using context-dependent embeddings. The Figure 6 controls are particularly informative, showing that removing context-specificity (averaged condition) or scrambling it (shuffled condition) both reduce performance. This confirms that the contextual advantage reflects genuine context-dependent neural representations.

3 Part 2: Novel Hypothesis

3.1 Hypothesis and Motivation

The original paper showed that both surprisal and contextual embeddings predict neural activity, but does not quantify their shared versus unique contributions. This leaves open how much the two signals overlap versus capture distinct aspects of neural processing. Surprisal is a scalar summary of prediction error and contextual embeddings encode rich semantic and syntactic structure. We hypothesized that these features capture partially independent neural signals. Zada et al. [4] recommends variance partitioning to measure unique contributions of different feature sets, which motivated our approach.

3.2 Methods

We performed variance partitioning at each time lag. For every electrode, we computed R^2 (squared encoding correlation, clamped at zero) under three ridge regression models ($\alpha = 1.0$, 2-fold cross-validation): surprisal only (1 feature), contextual embeddings only (50 features, layer 47, PCA), and a

combined model (51 features). We used ridge regression rather than OLS here because regularization penalizes model complexity, to account for differences in feature dimensionality across models (1 vs 50 vs 51). Without regularization, the 50-dimensional model would have more capacity to overfit, inflating the apparent dominance of embeddings. Unique variance was computed as $R^2_{\text{combined}} - R^2_{\text{other}}$, with shared variance as $R^2_{\text{surprisal}} + R^2_{\text{embeddings}} - R^2_{\text{combined}}$.

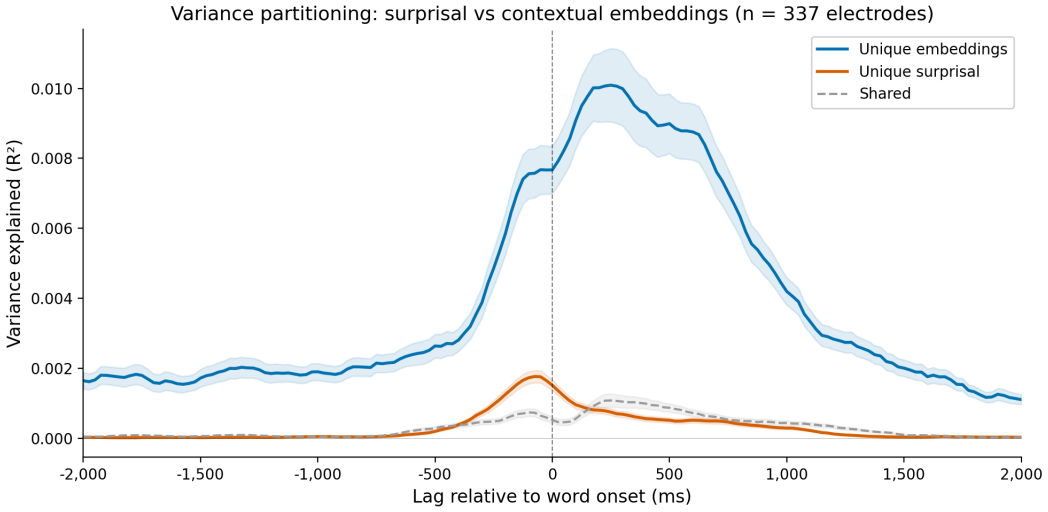

Figure 7: Time course of variance partitioning components: unique embeddings (blue), unique surprisal (orange), and shared variance (grey dashed). 337 electrodes.

3.3 Results and Discussion

At the peak lag (+200 ms), contextual embeddings accounted for 84.7% of total explained variance, shared variance was 8.5%, and surprisal contributed 6.8% uniquely. A Wilcoxon signed-rank test confirmed that unique embedding variance exceeded unique surprisal variance ($p = 1.52 \times 10^{-36}$), with the vast majority of electrodes showing embedding dominance.

The small shared component (8.5%) indicates that surprisal and contextual embeddings are largely complementary rather than redundant. This matches the theoretical distinction as well, since embeddings encode what the brain represents about a word in context, while surprisal encodes how unexpected that word was. The time courses of the variance components support this interpretation. Unique embedding variance peaks around word onset and persists, while unique surprisal variance emerges only post-onset, consistent with prediction error dynamics.

One issue is the dimensionality asymmetry, with 50 features for embeddings versus 1 for surprisal, which gives the embedding model more capacity even with regularization. A fairer comparison might use matched-dimensionality representations, for example extracting the top principal component of the embeddings for a direct 1-vs-1 comparison. That said, the presence of unique surprisal variance (6.8%) confirms that prediction error carries neural information not reducible to the contextual embedding. Overall, these results support the view that prediction and representation are dissociable computational principles in the brain's language system, and that a complete account of neural language processing requires both.

References

[1] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

[2] Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. Shared computational principles for

language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380, 2022.

[3] Jeffrey Pennington, Richard Socher, and Christopher D Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[4] Zaid Zada, Samuel A Nastase, Bobbi Aubrey, Itsik Jalon, Sebastian Michelmann, Haocheng Wang, et al. The "Podcast" ECoG dataset for modeling neural activity during natural language comprehension. *Scientific Data*, 12(1):1135, 2025.