Information-making processes in the speaker's brain drive human conversations forward

Ariel Goldstein[1,2♱], Haocheng Wang[3]*,Tom Sheffer[2]*, Mariano Schain[2]*, Zaid Zada[3], Leonard Niekerken[3], Bobbi Aubrey[3], Samuel A. Nastase[3], Harshvardhan Gazula[4], Colton Costo[4,5], Werner Doyle[6], Daniel Friedman[6], Sasha Devore[6], Patricia Dugan[6], Avinatan Hassidim[2], Michael Brenner[2,5], Yossi Matias[2], Orrin Devinsky[6], Adeen Flinker[6], Uri Hasson[3]

[1]Hebrew University, Jerusalem, Israel
[2]Google Research, Mountain View, CA, USA
[3]Princeton University, NJ, USA
[4]Massachusetts Institute of Technology, Cambridge, MA
[5]Harvard University, Cambridge, MA
[6]New York University School of Medicine, New York, NY, USA
*Equal contribution
♱Corresponding author: ariel.y.goldstein@mail.huji.ac.il

Abstract

A conversation following an overly predictable pattern is likely boring and uninformative; conversely, if it lacks structure, it is likely nonsensical. The delicate balance between predictability and surprise has been well studied using information theory during speech perception, focusing on how listeners predict upcoming words based on context and respond to unexpected information. However, less is known about how speakers' brains generate structured yet surprisingly informative speech. This study uses continuous electrocorticography (ECoG) recordings during free, 24/7 conversations to investigate the neural basis of speech production and comprehension. We employed large language models (Llama-2 and GPT-2) to calculate word probabilities based on context and categorized words into probable (top 30%) and improbable (bottom 30%) groups. We then extracted word embeddings from the LLMs and used encoding models to estimate the neural activity while producing or listening to probable and improbable words. Our findings indicate that before word-onset, the human brain functions in opposing, perhaps complementary, ways while listening and speaking. Results show that listeners exhibit increased neural encoding for predictable words before word onset, while speakers show increased encoding for surprising, improbable words. Speakers also show a lower speech production rate before articulating unexpected words, suggesting additional cognitive processes are involved in producing novel information. This indicates that human speech production includes information-making processes for generating informative words that are absent in language models, which primarily rely on statistical probabilities to generate contextually appropriate speech.

Introduction

Information theory mainly focuses on how listeners perceive information during natural communication (*1*–*3*). Entropy, in this framework, measures the uncertainty of an upcoming word given the preceding words (i.e., context). From the listener's perspective, predictable words provide little new information, while surprising words are informative as they deviate from what was expected based on the context and prior knowledge (*4*, *5*). The remarkable power of calibrating the predictions by minimizing cross-entropy (surprise) was recently unveiled by large language models (LLMs). LLMs acquire strong, almost human-like linguistic competence while relying on next-word prediction to robustly learn the statistical structure of natural language (*6*, *7*). Recent research has shown that, like LLMs, the inferior frontal gyrus (IFG) in the listeners' brains actively and implicitly predicts the next word before it is spoken while listening to real-life stories (*8*, *9*). After articulation, the IFG appears to calculate its surprise level (prediction error) and engage in further processing of the surprising (informative) words (*8*, *10*, *11*). This suggests that the listeners’ brains proactively seek new and surprising information while processing natural language.

However, what is information theory's role in generating information rather than perceiving it? At one extreme, the predictability of words in context may be sufficient to guide speech production, as suggested by LLMs' ability to produce coherent and meaningful text by solely selecting among the top most probable words in context. At the other extreme, the predictability level of words in context may have minimal relevance for speech production. After all, if the speaker intended to surprise their listeners, they could select randomly among words. For example, the word “cow,” while surprising in the context of this paragraph, provides very little information to the readers. Therefore, the speaker may not aim to surprise but to convey meaningful information. In this case, during speech production, the internal predictive process may be replaced with a process in which the speaker carefully chooses the words they are about to say, independently of how probable they are from the listeners’ perspective. Our findings support an intermediate stance where probability and information-making shape spontaneous speech production in everyday conversations.

The neural basis of spontaneous speech production is one of the least studied and least understood aspects of human cognition. Most existing research on speech production used highly controlled experimental conditions, focusing on articulating predetermined words or sentences rather than exploring the complexities of generating speech in everyday contexts (*12*–*14*). While numerous studies have examined how the listener’s brain processes probable (predictable) and improbable (surprising) words embedded in natural contexts, little is known about the underlying neural processes that support spontaneous speech production. To investigate how the brain processes natural language in real-life situations, we gathered a unique 24/7 dataset of continuous electrocorticography (ECoG) and spontaneous conversations throughout the patients' day-to-week-long stays at the NYU Medical School's epilepsy unit (*15*). In our setup, patients are free to say whatever they want, whenever they want; each conversation has its unique context and purpose. Thus, for the first time, we can study the neural basis of spontaneous speech production (information-making) and speech comprehension (information-seeking) within the same set of participants. This ambitious effort resulted in a uniquely large ECoG dataset of natural conversations: four patients recorded during free conversations, yielding approximately 50 hours (230,238 words) of neural recordings during speech production and 50 hours (289,971 words) during speech comprehension.

LLMs are powerful statistical models for analyzing our 24/7 conversational data. Thus, we used state-of-the-art LLMs (Llama-2 and GPT-2) to assign a probability to each word in each conversation as a function of all prior words (context). Next, we divided all words into probable (top 30%) and improbable (bottom 30%) words. We constructed electrode-wise encoding models to estimate a linear mapping from the word embeddings in each LLM to the neural activity for each word during speech production and comprehension. This allowed us to directly compare neural processing in the same participants while listening to or producing probable and improbable words in natural, real-life conversations.

Our findings indicate that before word-onset, the language areas function in opposing, perhaps complementary, ways while listening and speaking. In listeners' IFG, we reproduced our previous finding of enhanced pre-word-onset encoding for probable versus improbable words in speech comprehension (*8*, *9*). Conversely, in speakers' IFG, we found, for the first time, enhanced pre-word-onset encoding for improbable versus probable words. The results remained strong and clear even when we narrowed down the analysis to a shared set of words that were unlikely in one context and likely in another. This confirms that the observed effect can be decoupled from the word frequency effect that previous studies have documented. Behaviorally, all speakers exhibited lower speech rates before uttering improbable words. These findings suggest that additional cognitive processes are involved in generating surprising words in the speaker's brain, processes that are not required for generating probabilistic speech in LLMs. Our findings challenge the notion that LLMs suffice for capturing the complexities of human language generation. In contrast, we propose a nuanced perspective on conversations, in which the speaker's brain aims to incorporate informative (surprising) words into a structured (probable) context while the listener's brain attempts to identify and learn from its failures to predict the next probable word in context.

Results

Our 24/7 conversation data consists of half a million words recorded during 100 hours of real-life conversations between our four ECoG patients and their surroundings in the hospital room. The conversations are spontaneous and cover various real-life topics, including discussions between the patients and medical staff and personal conversations about family, friendships, sports, and politics. We used LLMs (Llama-2 and GPT-2) to predict the next word based on all prior words (context) in each conversation.

Our analysis of recorded natural conversations found that approximately 25% of the half-million words spoken were entirely predictable using a Llama-2 top-1 prediction (and 23% for GPT-2; Supp. Fig. 1). The top 2 predictions accounted for 34% of the total words (and 31% for GPT-2). Moreover, Llama-2 accurately predicted around 70% of all words by focusing on a small set of roughly the top 22 most probable words in a given context (and 34 for GPT-2). Given the low chance of accurately predicting a word from a lexicon containing tens of thousands of words, this highlights the highly structured nature of natural conversations. However, certain aspects of natural conversations are always harder to predict (Supp. Fig. 1). It would take over 50 predictions (Llama-2: 54; GPT-2: 83) to accurately predict 80% of the words, hundreds (Llama-2: 167; GPT-2: 300) to predict 90% of the words, and thousands of predictions to account for all words. We divided the words into probable (top 30%) and improbable (bottom 30%) words using LLMs' prediction accuracy levels and LLMs' confidence levels (Supp. Table 1).

Next, we extracted non-contextual word embeddings from the LLMs and used encoding models to estimate the neural activity while producing or listening to probable and improbable words. The superior spatiotemporal resolution and signal-to-noise ratio (SNR) of our 24/7 ECoG recordings enable us to focus this paper on the neural processes before word onset in the same individuals during speech production and comprehension. This is as opposed to prior research, which, due to the limited SNR and spatiotemporal resolution of EEG, MEG, and fMRI methods, has primarily focused on assessing the post-word-onset surprise effect, such as the P300 and N400 markers (*16*, *17*).

During speech comprehension (listening), we observed enhanced encoding for upcoming probable words (top 30% probability based on accuracy) compared to improbable words (bottom 30%) hundreds of milliseconds before word onset (Fig. 1A). This finding replicates our recent discovery of enhanced encoding for probable words before word onset during passive listening to a podcast (*8*), using new data from spontaneous conversations. In addition, we observed enhanced encoding around 300 ~ 400 ms after word onset in the listeners' brains for the improbable (surprising) words (Fig. 1A).

During speech production (speaking), the same participants showed a reversed effect, with pre-word-onset enhanced encoding for improbable (surprising) words over probable words (Fig. 1B). The enhanced encoding for improbable words in the speaker's brain was apparent across multiple brain regions (Fig. 1B, see also additional ROIs in Supp. Fig. 2). The contrast between enhanced encoding for improbable words in the speaker's brain and enhanced encoding for probable words in the listener's brain appears robust. First, it was replicated when we limited the analysis to content words such as nouns, verbs, adjectives, and adverbs while excluding all function words (Supp. Fig. 3). Secondly, it was replicated using only a shared list of words present in both probable and improbable groups (Supp. Fig. 4). This suggests that the effect can be independent of the words' frequency base in natural language. Thirdly, it was replicated when we relied on the models' confidence level rather than its accuracy level, controlling for the number of probable and improbable words (Supp. Fig. 5). Finally, the results are robust across different language models, as we replicated the effect using GPT-2 predictions and embeddings (Supp. Fig. 6).

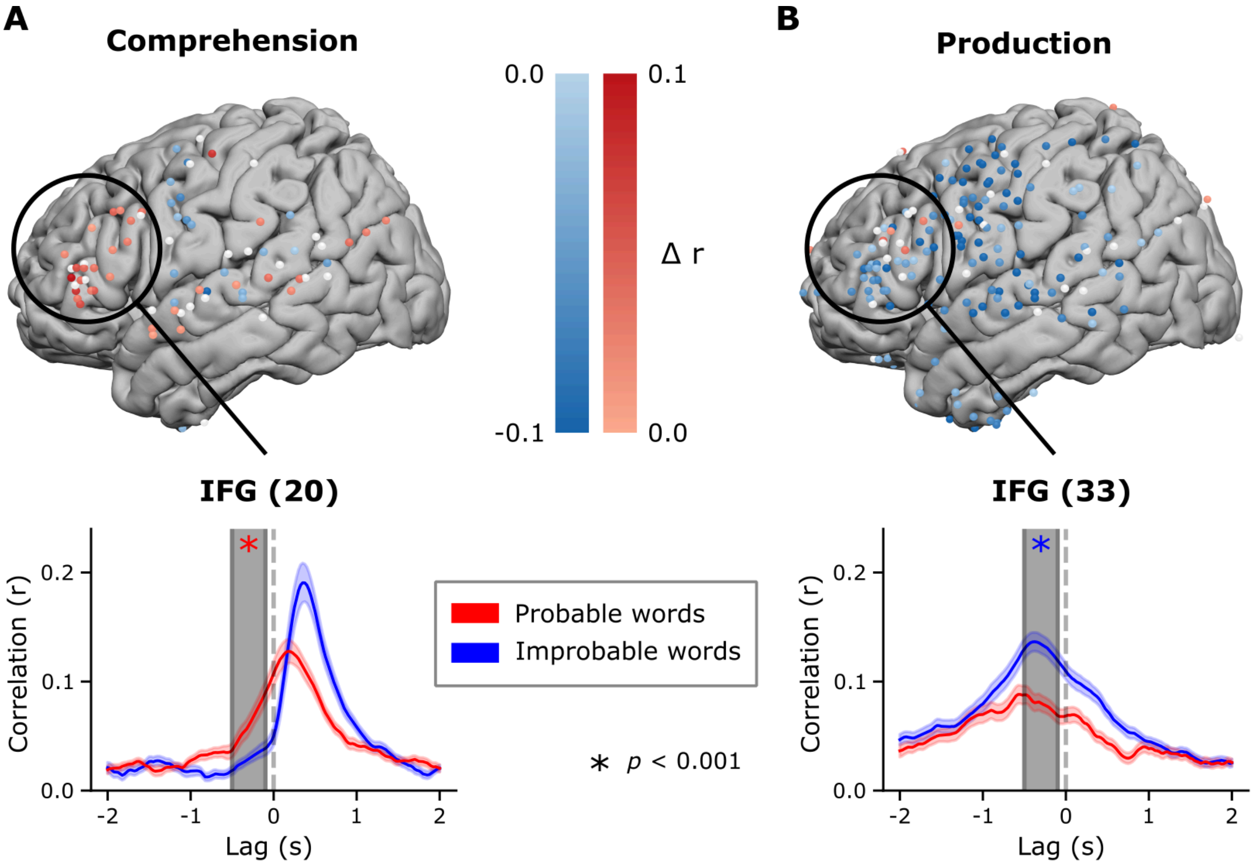


Figure 1. Encoding Results in the Inferior Frontal Gyrus (IFG) for Probable and Improbable Words. A. During speech comprehension (listening), we noticed enhanced encoding for probable words compared to improbable words around 100 - 500 ms (gray bar) before the words were spoken. This processing was based on word embeddings extracted from Llama-2. **B.** We observed the opposite effect during speech production (speaking), with enhanced encoding for improbable words over probable words about 100 - 500 ms (gray bar) before the word was spoken. Brain maps: The color scales indicate encoding differences between probable and improbable words averaged across lags (-500 to -100 ms). Red (blue) electrodes showed significantly increased encoding for probable (improbable) words ($q <$ 0.001, FDR corrected). White indicates electrodes with no statistically different encoding for probable versus improbable words.

Behaviourally, speakers slow their speech rate and pause for an additional 100 - 150 ms ($p <$ 0.001, for full statistical details see Supp. Table 2) before articulating improbable or surprising words (Fig. 2A). This pattern was observed in all four participants (S1-S4), as well as in the analysis of all other speakers who participated in our conversations, for whom brain responses were not recorded (Fig. 2). The pattern was independent of the words' frequency as the results hold when introducing words’ frequency (*18*) as a covariate (*p* < 0.001) and when the analysis was restricted to a shared set of words across the probable and improbable word lists (*p* < 0.001, Fig. 2B, Supp. Table 2).

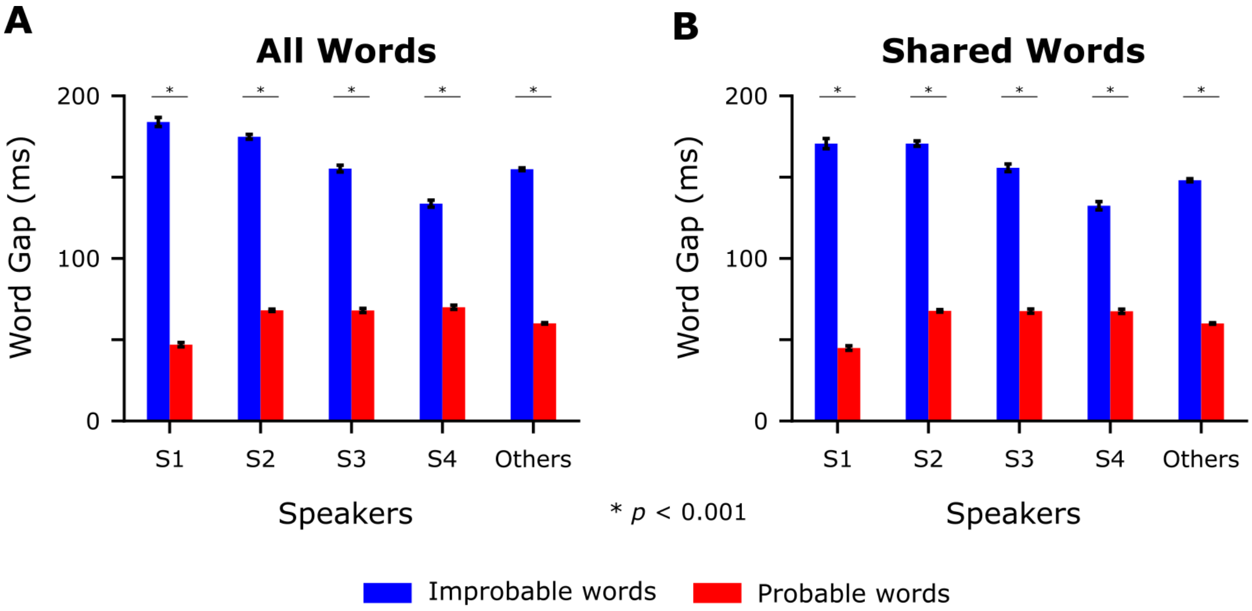


Figure 2. Behavior Temporal Gap between the Offset of the Previous Word and Onset of the Current Word. A. It takes about 100 - 150 ms longer for each speaker (S1-S4) to start articulating improbable words. This was also evident when examining the data of other speakers in the room, for whom brain responses were not recorded. **B.** The pause before word onset for improbable words was consistent, even when the analysis was limited to a shared set of words across both improbable and probable word lists. This suggests that the pause was independent of the word's frequency in the natural language.

Discussion

Conversations that follow a predictable pattern can be uninformative, while those lacking structure seem nonsensical. Like the yin and yang, a pair of equal opposites that attract and complement each other, conversations must be generated in a process that adheres to the shared statistical structure of language while simultaneously compelled to violate it in informative ways. In our study, for the first time, we have discovered evidence of information-making processes in the speaker's brain before articulating words that complement the information-seeking processes in the listeners' brains.

Conversations are not possible without a shared alignment of context among speakers (*19*). After all, words can only be meaningfully surprising against a backdrop of shared, structured, and predictable context. This is why speakers tend to establish a common ground and shared context at the beginning of each utterance before conveying new, surprising information (*20*, *21*). Indeed, in our analysis, we found robust encoding before word onset in the speaker's brain for probable words within all language areas (red line in Supp. Fig. 2). The shared context between the speaker and their audience provides the necessary background for conveying new information. Furthermore, the efficiency and speed in producing and understanding speech heavily rely on a shared and well-structured linguistic system (*22*, *23*). Studies on intercultural communication underscore the importance of shared linguistic conventions for mutual understanding and effective interaction (*24*). Moreover, it was recently demonstrated that during face-to-face conversations, the neural activity of the speaker and the listener become aligned to a set of shared linguistic features that can be captured by an LLM (*25*).

Fully predictable conversations, however, become uninformative. Therefore, there is little point in conversing if the speaker fails to convey new and surprising information to their listeners. But how does the speaker generate informative conversations? Surprisingly, information theory provides little insight into information-making processes. The recent development of LLMs illustrates this. LLMs learn language by relying on information-seeking procedures like next-word prediction as the model processes incoming human conversations. However, there's no need for additional processes in LLMs to generate surprising yet meaningful words during speech production. After all, the same process of selecting the next word from the learned context-dependent probability distribution (using a temperature parameter) is used for all spoken words. The lack of additional information-making processes during speech production may explain why LLMs deteriorate when trained with text generated by other LLMs rather than humans (29). In such cases, the generated probabilistic text becomes more predictable and less informative over time, leading to a rapid collapse of the language model. The intuition that something essential is missing in how LLMs generate language is supported by our findings of novel information-making processes in the speaker's brain that are not present in LLMs. These processes enable humans to innovate and intentionally deviate from the statistical structure of language.

Behaviourally, our analysis found that speech slows by about 100 - 150 ms before articulating words that are unlikely or hard to predict. Previous research has indicated that speakers take longer to start articulating rare (infrequent) words (*26*, *27*). Such a process was attributed to difficulty retrieving from memory or planning an articulatory motor sequence for infrequent words. Our findings indicate that our effect is context-specific rather than word-frequency-specific, as the pause in articulation can be longer for the exact same words spoken in unpredictable versus predictable contexts. Furthermore, in addition to the increased gap before the articulation of improbable words, it was established that the duration for articulating them also slowed down (*28*). The slow-down in speech rate provides a behavioral marker for additional cognitive processes in the speaker's brain that take place before conveying new difficult-to-predict information.

Neurally, our analysis found evidence for enhanced neural encoding in the speaker's brain for improbable words over probable words, suggesting additional information-making neural processes during the spontaneous production of surprising words (Fig. 1B). The enhanced encoding for improbable words before word-onset is the opposite to that of enhanced encoding for probable words, before word-onset, found in the listener's brain (Fig. 1A). Moreover, while next word prediction in the listener's brain was localized to the IFG, the speaker's brain demonstrates enhanced encoding for improbable words across various areas, including the IFG, STG, angular gyrus, and precentral motor cortex (Fig. 1B and Supp. Fig. 2).

Finally, the enhanced encoding for improbable words before speaking in the speaker's brain is mirrored by the enhanced encoding for improbable words after speech articulation in the listener's brain (Fig. 1). This suggests that both the speaker (before word-onset) and the listener (after word-onset) home-in on the informative and unpredictable words in each conversation.

This study has several limitations. First, the nature of the neural processes in the speaker’s brain associated with choosing improbable yet informative words is not yet defined. While we observe improved encoding of improbable words in the speaker's brain, we know very little about the underlying neural processes that guide the selection of informative words. The extensive focus of previous research on information-seeking in listeners leaves a theoretical gap in our understanding of the neural processes of information-making in the speaker’s brain, which is likely to be linked to the human capacity to think and innovate. Furthermore, while information-seeking and entropy are fundamental for understanding speech comprehension, entropy (surprise) gives us only a narrow window into speech production. Random words, by their nature, are unpredictable and, therefore, have high entropy. They introduce surprise but do not convey meaningful information because they lack context or relevance to the message. Thus, it is clear that apart from the element of surprise, other computations must be involved in the process of information-making.

Second, to determine the level of surprise for each word in the conversation, we depend on the exceptional capacity of LLMs to assign a probability to each word in any given conversation. Previous research has shown good agreement between people and LLMs' capacity to predict the next work in context (*8*). However, the ability to assess the level of surprise using LLMs is likely conservative because it lacks access to the specific history and shared knowledge among our speakers. For example, a family member may know that the patient loves frozen bananas, even though it may be a rare and improbable utterance for LLMs. The lack of access to the unique shared knowledge among speakers works against us, leading to a noisier assessment of the true level of surprise of the bottom 30% of the words. Indeed, a recent study showed that fine-tuning LLMs to better align with listeners' prior knowledge could improve the alignment of the LLM’s embeddings and the neural responses of subjects listening (*29*). Given the noisier estimate of how surprising they are in a given context, the enhanced encoding of improbable words in the speaker's brain demonstrates the strength of our results.

To conclude, the novel 24/7 ECoG recordings of natural conversation provide a new window to information-making processes in the speakers' brains, complementing the information-seeking processes in the listeners’ brains. These generative, information-making processes have been overlooked in information theory, neuroscience, and psychology due to excessive focus on speech comprehension processes. They also seem absent in LLMs, which rely solely on probabilistic speech to generate conversations. However, these information-making processes may be the key to understanding our ability to use natural language to think, innovate, and reinvent ourselves and our culture.

Methods

Preprocessing the speech recordings

We developed a semi-automated pipeline for preprocessing datasets consisting of four main steps. First, we de-identified speech recordings by manually censoring sensitive information to comply with HIPAA regulations. Second, we used a human-in-the-loop process with Mechanical Turk transcribers to accurately transcribe the noisy, multi-speaker audio. Third, we aligned text transcripts with audio recordings using the Penn Forced Aligner and manual adjustments for precise word-level timestamps. Finally, we synchronized speech with neural activity by recording audio through ECoG channels, achieving 20-millisecond accuracy for aligning neural signals with conversational transcripts. For a full description of the procedure, see (*15*).

Preprocessing the ECoG recordings

We developed a semi-automated analysis pipeline to identify and remove corrupted data segments (e.g., due to seizures or loose wires) and mitigate other noise sources using FFT, ICA, and de-spiking methods (*30*). Neural signals were bandpassed (75–200 Hz), and power envelopes were computed as proxies for local neural firing rates (*31*). The signals were z-scored, smoothed with a 50 ms Hamming kernel, and trimmed to avoid edge effects. Custom preprocessing scripts in MATLAB 2019a (MathWorks) were used for these steps. For a full description of the procedure, see (*15*).

Prediction and embedding extraction

We extracted contextualized predictions and static word embeddings from GPT-2 (gpt2-xl, 48 layers) and Llama-2 (Llama-2-7b, 32 layers). We used the pre-trained version of the model implemented in the Hugging Face environment (*32*). We first converted the words from the raw transcript (including punctuation and capitalization) to whole words or sub-word tokens. We used a sliding window of 32 tokens (results were also replicated for 1024 tokens), moving one token at a time to extract the embedding for the final token in the sequence. Encoding these tokens into integer labels, we fed them into the model, and in return, we received the activations at each layer in the network (also known as a hidden state). For the predictions, we extracted the logits from the model for the second-to-last token, which was utilized by the model to predict the last token. For the embeddings, we extracted the activations for the final token in the sequence from the 0-th layer of the model before any attention modules. For tokenized words to be divided into several tokens, we take the prediction values of the first token and average the embeddings across several tokens. With embeddings for each word in the raw transcript, we aligned this list with our spoken-word transcript that did not include punctuation, thus retaining only full words.

Electrode-wise encoding

We used linear regression to estimate encoding models for each electrode and lag relative to word onset, mapping our static embeddings onto neural activity. The neural signal was averaged across a 200 ms window at each lag (25 ms increments). Using ten-fold cross-validation, we trained models to predict neural signal magnitudes based on GPT-2 or

Llama-2 embeddings. Embeddings were standardized and reduced to 200 dimensions via PCA (we replicated results using PCA to 50 dimensions and ridge regression). Regression weights were estimated using ordinary least-squares regression and applied to the test set. Pearson correlation assessed model performance across 161 lags from -2,000 to 2,000 ms in 25-ms increments. For a full description of the procedure, see (*8*).

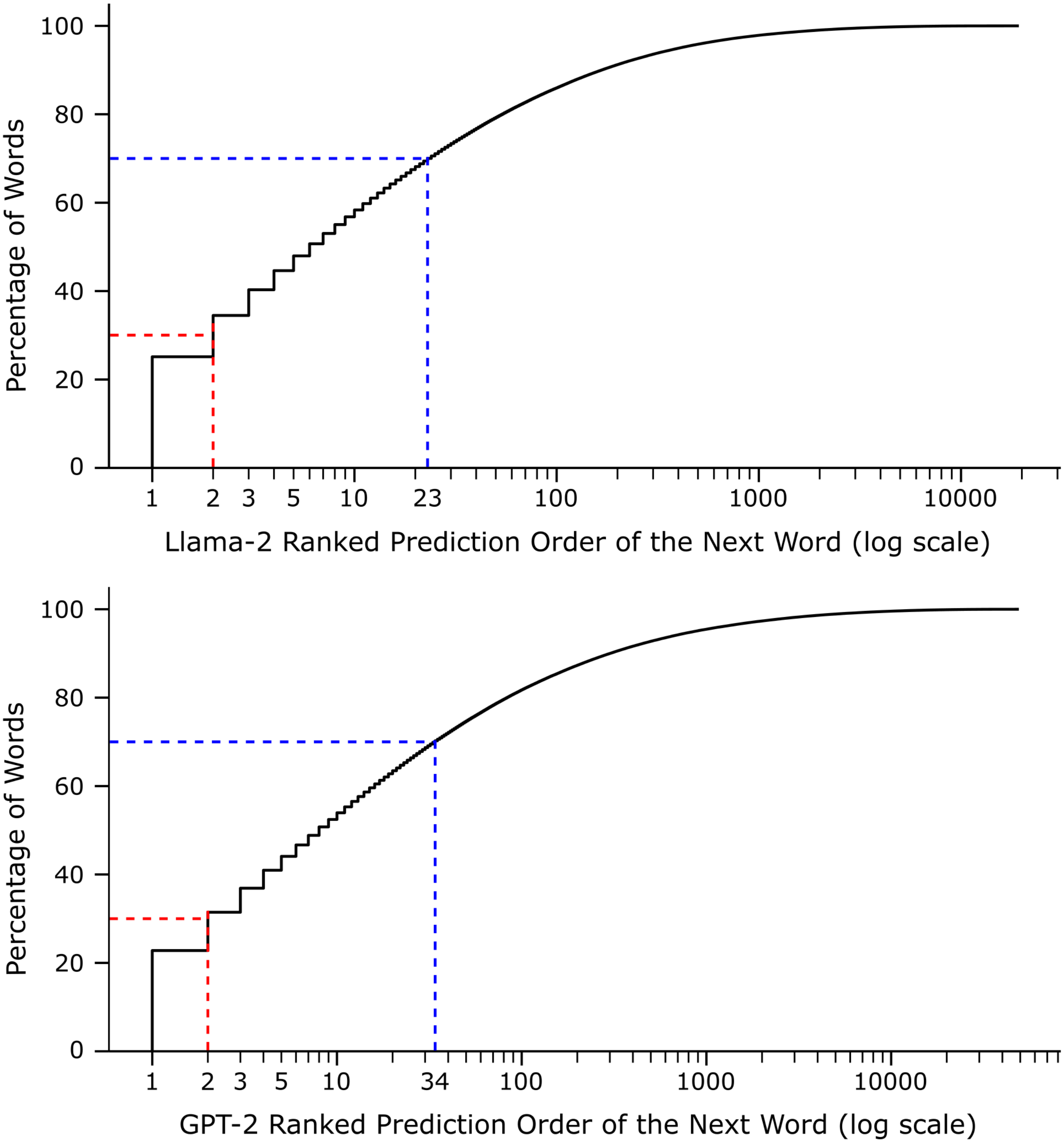
Electrode selection

A randomization method was employed to determine significant electrodes that were selective for semantic information. Each iteration involved randomly shifting embeddings (GloVe) assigned to predicted signals, breaking their connection with brain signals while maintaining their order without rolling over within the context window. The encoding procedure was then conducted for each electrode using the misaligned words, repeated 1,000 times. The score for each electrode was calculated by the range between the maximum and minimum values across 161 lags. From these, the highest value for each patient across all electrodes was recorded, forming a distribution of 1,000 maximum values per patient. The significance of electrodes was assessed by comparing the original encoding model's range to this distribution, calculating a p-value for each electrode. This tested the hypothesis of no systematic relationship between brain signals and word embeddings, resulting in family-wise error rate corrected p-values. Electrodes with p-values under 0.01 were deemed significant. For a full description of the procedure, see (*8*).

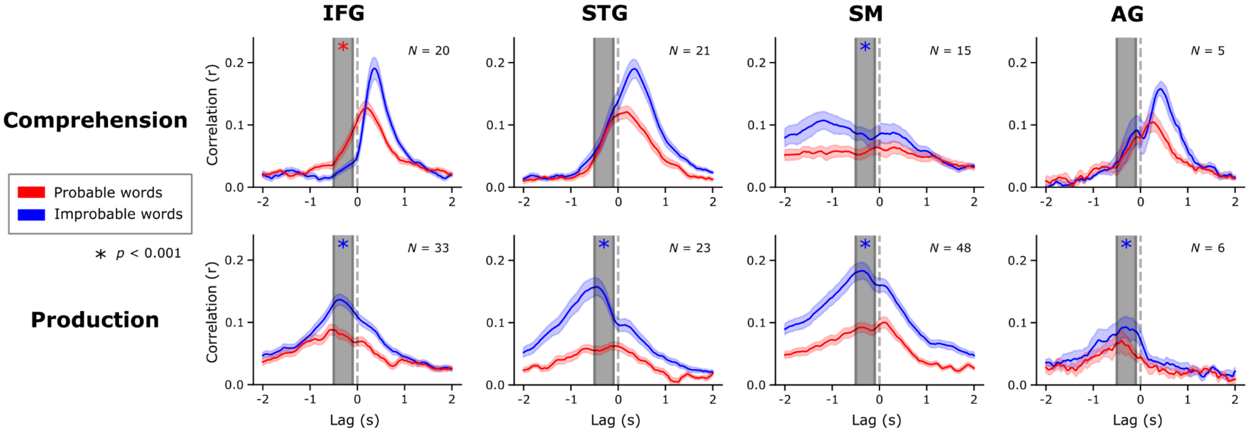
Significance test for encoding difference at the ROI level

To test for significant differences in encoding performance between probable and improbable word conditions in 17 given lags (-500 ms to -100 ms) for a specific ROI, we used a paired-sample permutation procedure: in each permutation, we randomly shuffled the labels (probable/improbable) of all observations (correlation encoding) for both conditions, and we computed that difference of the averages. A *p*-value was computed as the percentile of the non-permuted difference between the averaged correlation values for the probable and improbable words over the electrodes and lags relative to the null distribution. *P*-values less than 0.0005 (significance of 0.001 for the two-sided test) were considered significant. We used a similar paired-sample permutation procedure to test for significance for specific electrodes with samples from the 17 given lags. FDR correction was applied to correct for multiple electrodes.

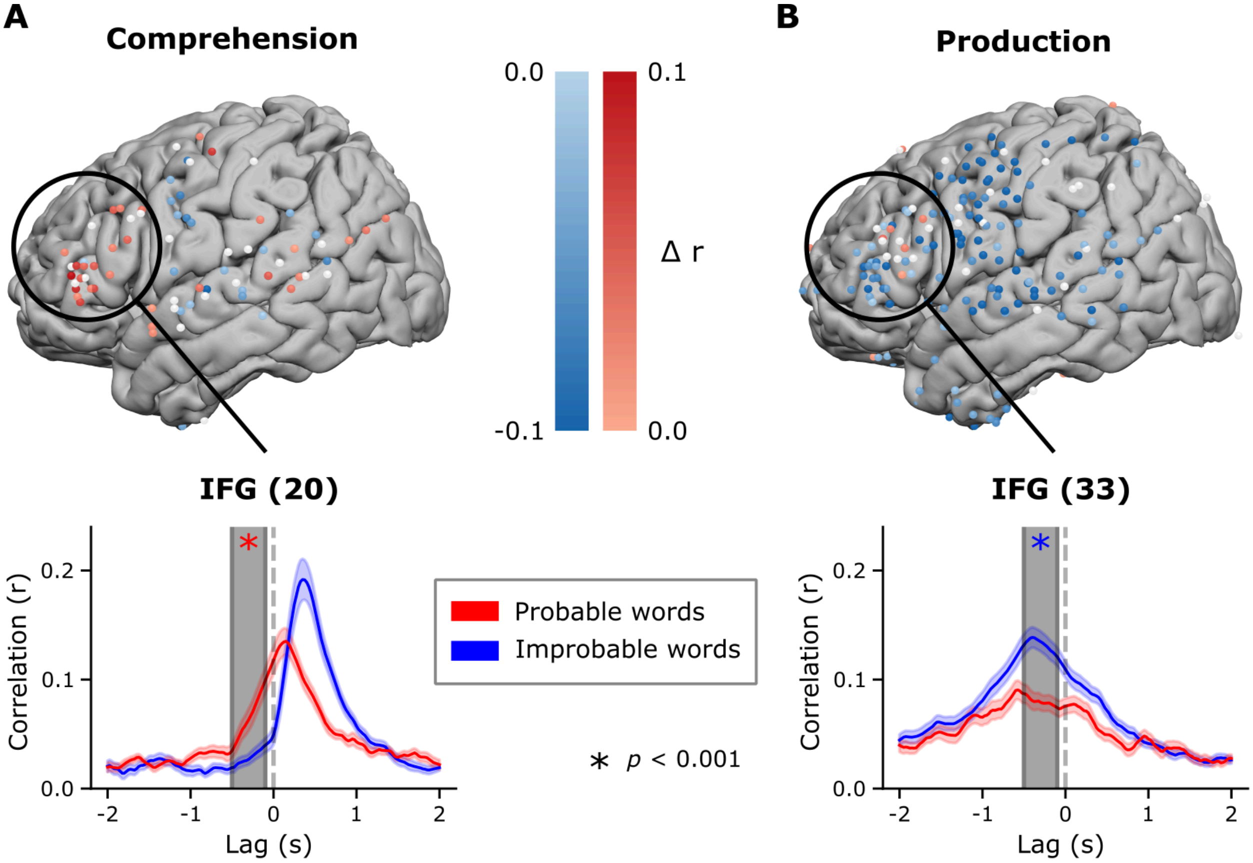
Supplementary Figures



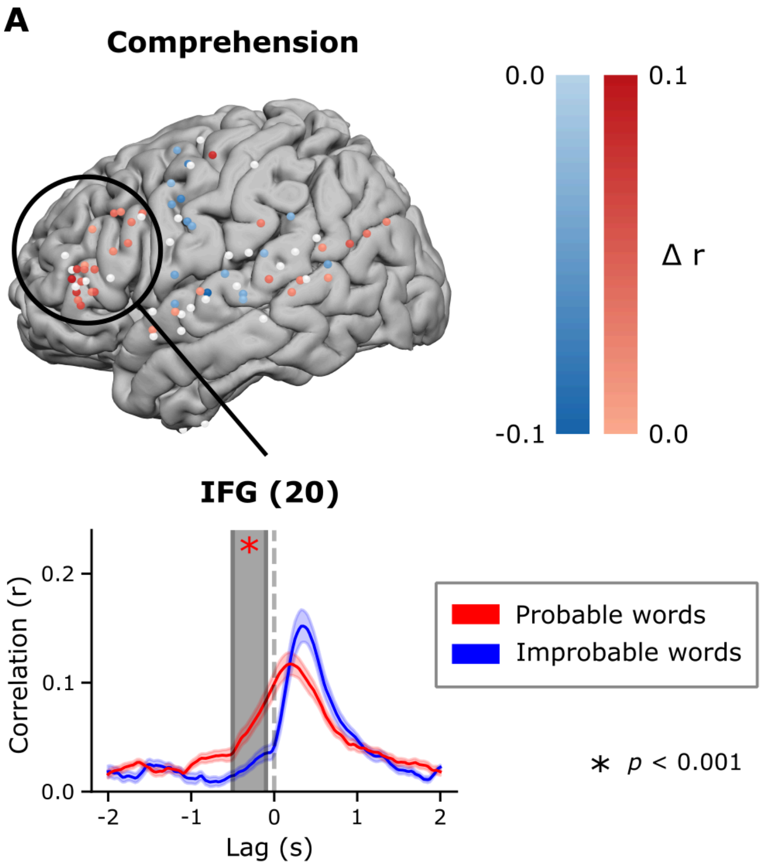
Supplementary Figure 1. Accumulated ranked-order predictions for upcoming words as predicted by Llama-2 and GPT-2. We extracted each next word's ranked probability according to Llama-2 (up) and GPT-2 (bottom) context-based predictions. The rank order is represented on a logarithmic scale. LLMs successfully predicted more than 25% of the words (top-1). Around 23/34 predictions were necessary to accurately forecast 70% of the words, while tens to hundreds of predictions were needed to predict the bottom 30%.

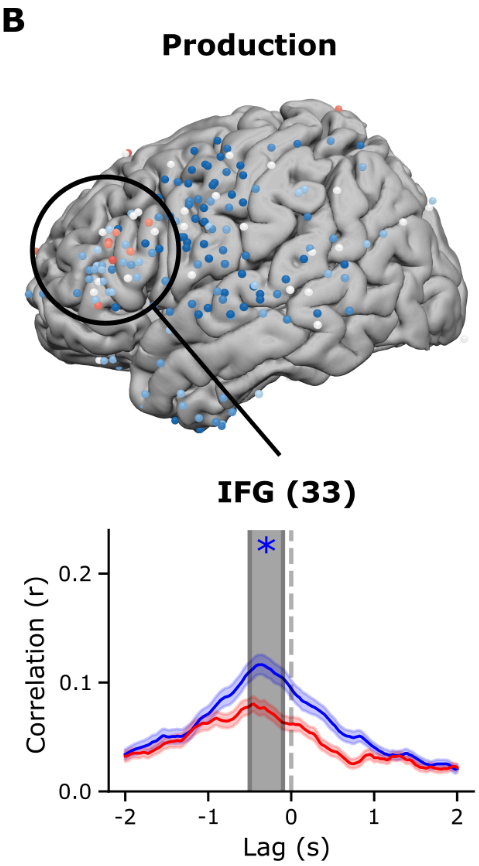


Supplementary Figure 2. Encoding Results for Probable and Improbable Words in Different ROIs. The listener's brain showed enhanced pre-word-onset encoding of probable words in the IFG, while the speaker's brain exhibited widespread enhanced pre-word-onset encoding of improbable words across several language areas.

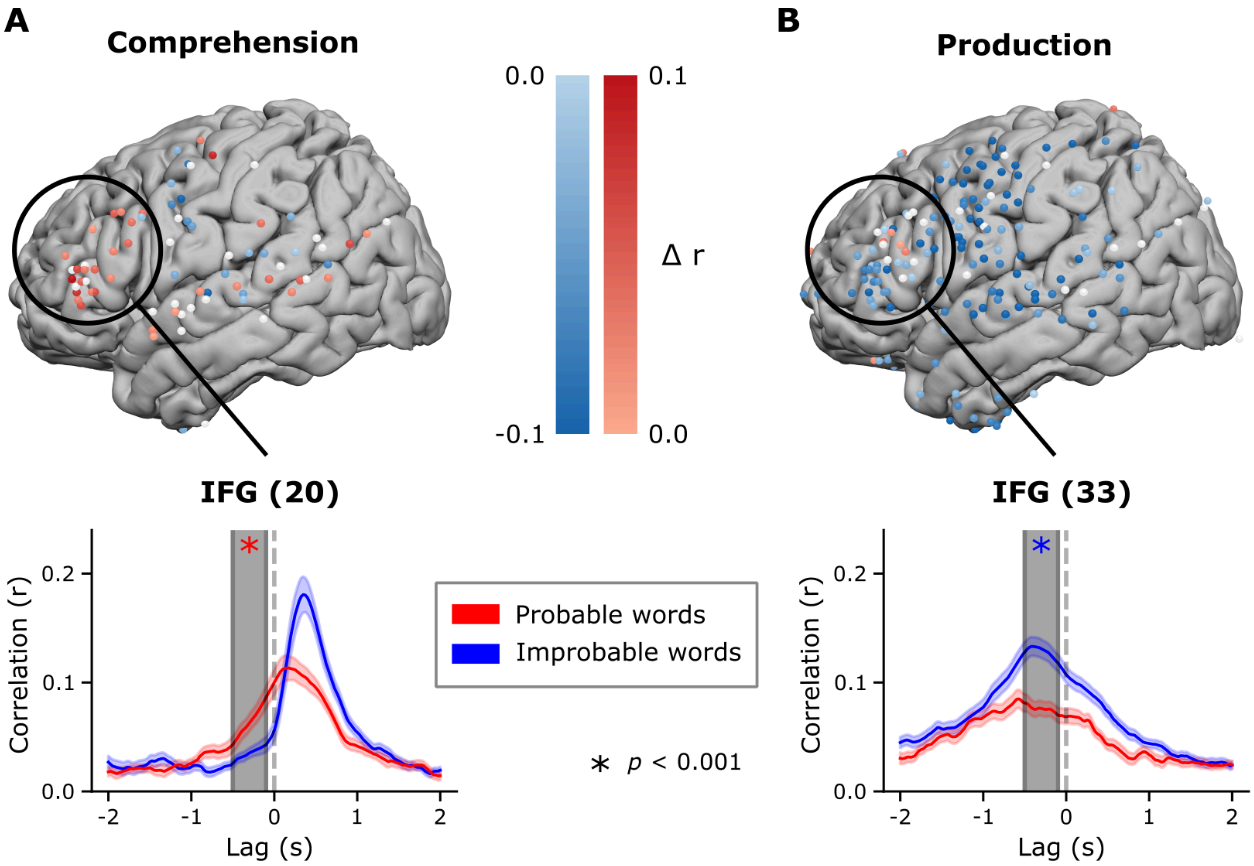


Supplementary Figure 3. Using content words to encode probable and improbable words. Similar results to those in Fig. 1 were achieved while restricting the encoding analyses to content words (i.e., nouns, verbs, adjectives, and adverbs, N = 306,681). This demonstrates that highly predictable function words do not drive the observed effect.

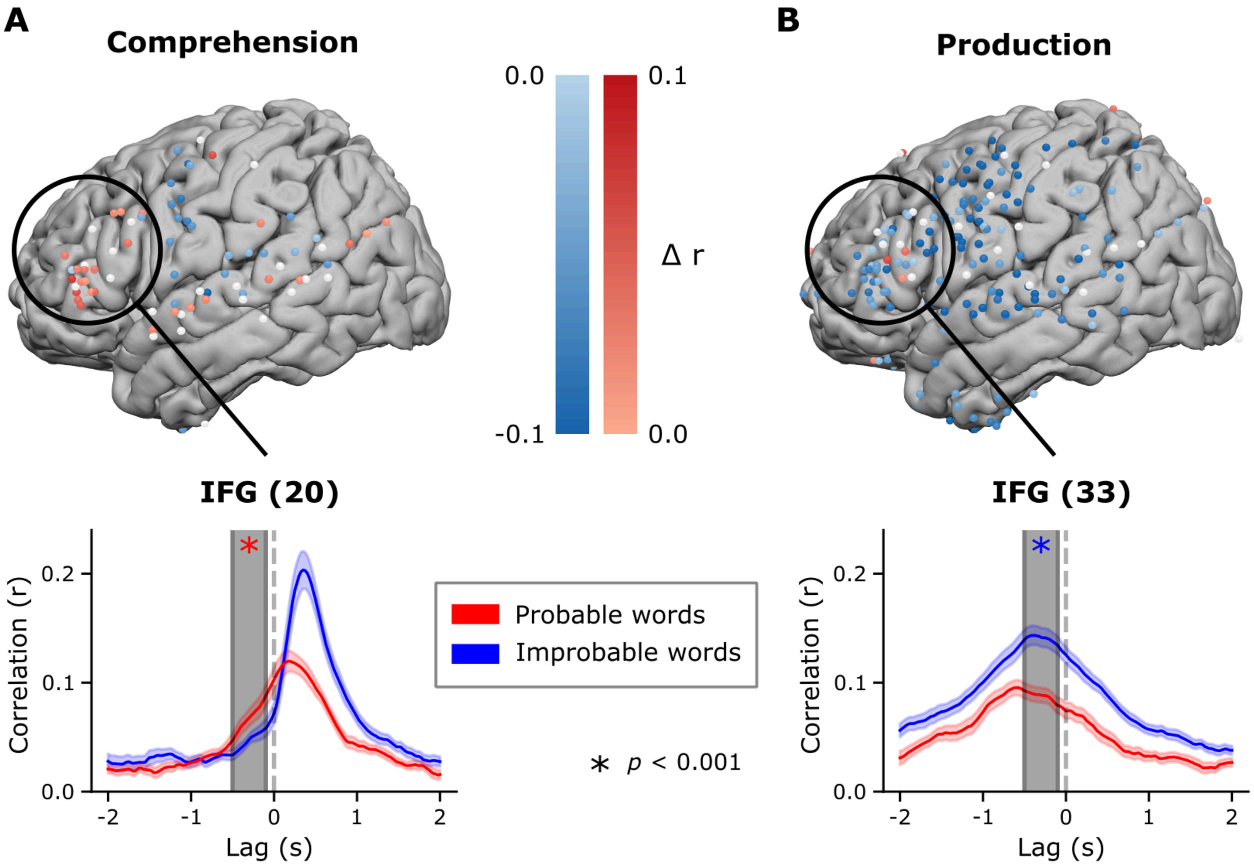




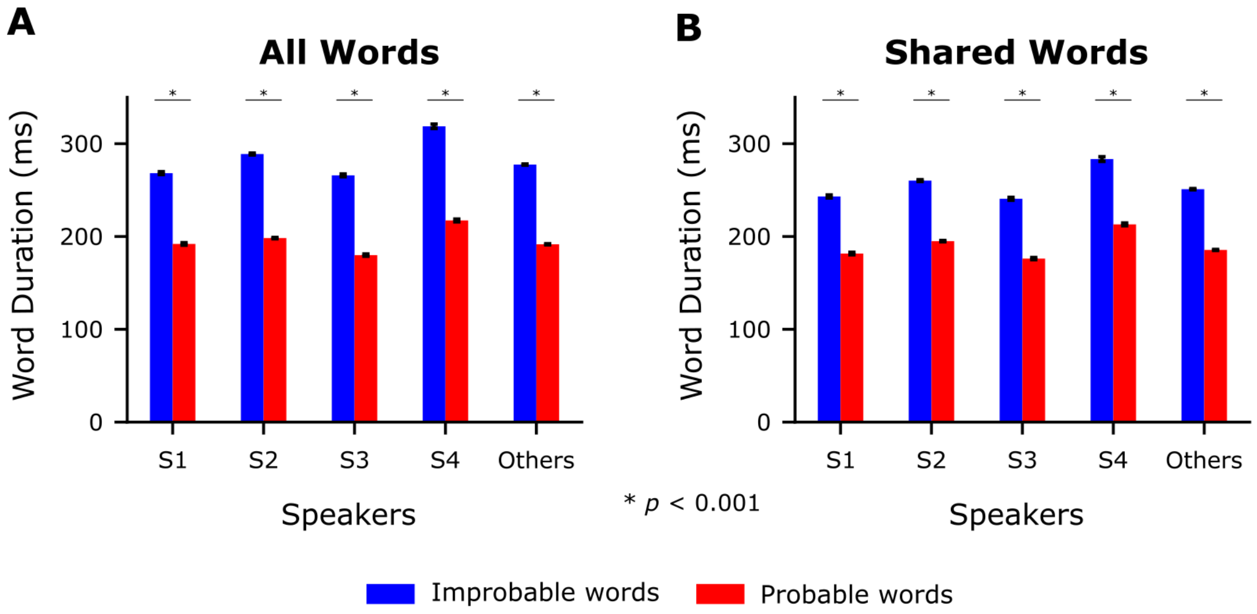
Supplementary Figure 4: Utilizing a shared set of words to encode probable and improbable words. Similar results to those in Figure 1 and Supplementary Figure 3 were achieved using a shared set of words, which were predictable in one context and unpredictable in another. This demonstrates that the observed effect can be decoupled from the word frequency effect that previous studies have documented.



Supplementary Figure 5. Using the model's confidence level to encode probable and improbable words. Similar results to those in Figure 1 and Supplementary Figures 3,4 were achieved using Llama-2's internal confidence level. This demonstrates that the observed effect can be replicated when we rely on the model's internal confidence rather than the model's success in predicting the next word (accuracy level).

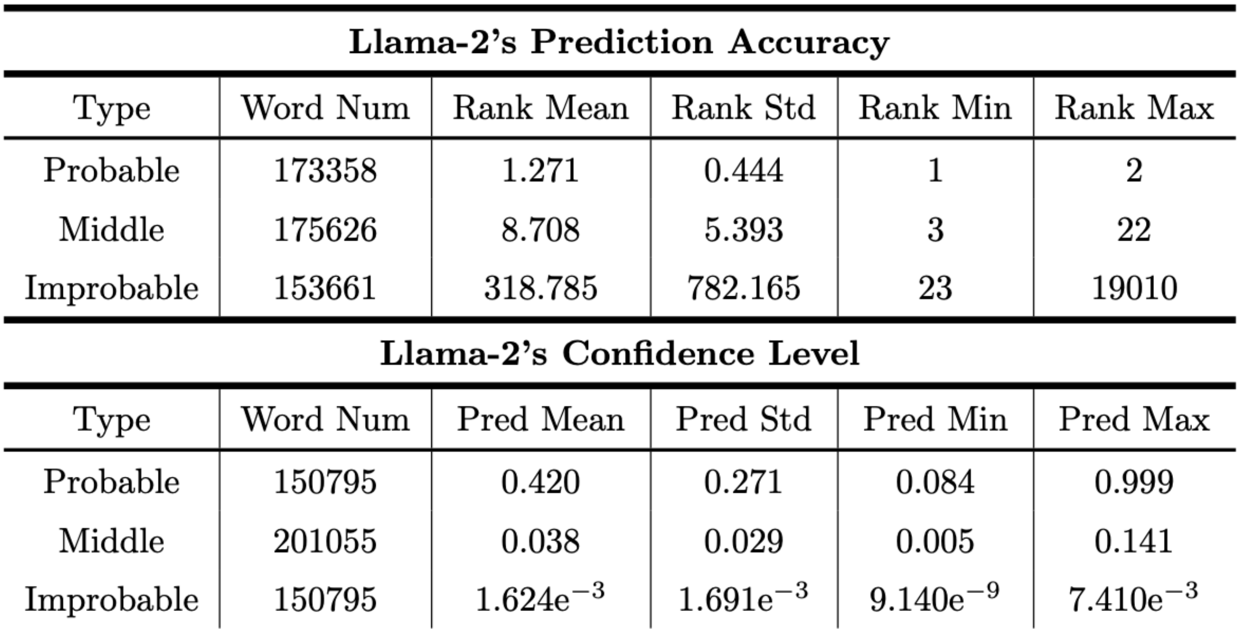


Supplementary Figure 6. GPT-2's predictions and embeddings are used to encode probable and improbable words. Similar results to those in Figure 1 and Supplementary Figures 3,4,5 were achieved using predictions and embeddings from GPT-2 instead of Llama-2. This demonstrates that our results can be reproduced using other LLMs.



Supplementary Figure 7. Behavior Temporal Duration between the Onset and Offset of the Current Word. In addition to the delay (silence) before speaking unlikely words (shown as the word gap effect in Fig. 2), it also took longer to pronounce unlikely words (A), even when we restricted the analysis to the same set of words which were probable in one context and improbable in another (B).

Supplementary Tables



Llama-2's Prediction Accuracy

Type	Word Num	Rank Mean	Rank Std	Rank Min	Rank Max
Probable	173358	1.271	0.444	1	2
Middle	175626	8.708	5.393	3	22
Improbable	153661	318.785	782.165	23	19010

Llama-2's Confidence Level

Type	Word Num	Pred Mean	Pred Std	Pred Min	Pred Max
Probable	150795	0.420	0.271	0.084	0.999
Middle	201055	0.038	0.029	0.005	0.141
Improbable	150795	$1.624e^{-3}$	$1.691e^{-3}$	$9.140e^{-9}$	$7.410e^{-3}$

Supplementary Table 1. Statistics of Words Divided into Probable (top 30%), Improbable (bottom 30%), and Middle (middle 40%) using Llama-2's prediction accuracy (top table) and confidence

levels (bottom table). Rank is the ranked prediction order of the next word, ranging from 1 to 32,000 (vocab size for Llama-2). Pred is the prediction probability of the next word, ranging from 0 to 1.

Word Gap (ms) for Probable/Improbable Words

Statistics	**Speaker 1**	**Speaker 2**	**Speaker 3**	**Speaker 4**	**Other Speakers**
Probable Mean	46.934	68.067	67.981	69.937	60.039
Probable Std	126.885	135.462	149.617	169.161	133.546
Improbable Mean	183.928	174.839	155.278	133.684	154.871
Improbable Std	250.426	228.850	231.330	228.029	227.956
Independent t-test	$t(16940) = 45.483$	$t(57756) = 70.212$	$t(26082) = 36.522$	$t(28302) = 26.968$	$t(157602) = 102.735$
	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$
ANCOVA	$F_{1,11555} = 851.521$	$F_{1,47734} = 3336.353$	$F_{1,21655} = 538.538$	$F_{1,22312} = 218.568$	$F_{1,128483} = 4306.434$
	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$

Word Gap (ms) for Shared Probable/Improbable Words

Statistics	**Speaker 1**	**Speaker 2**	**Speaker 3**	**Speaker 4**	**Other Speakers**
Probable Mean	44.865	67.728	67.575	67.481	59.973
Probable Std	124.782	134.938	149.192	166.274	133.824
Improbable Mean	170.646	170.640	155.722	132.408	148.079
Improbable Std	240.626	226.462	231.683	227.087	222.123
Independent t-test	$t(14133) = 40.432$	$t(52571) = 65.276$	$t(23098) = 35.123$	$t(24308) = 25.224$	$t(136314) = 91.368$
	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$

Supplementary Table 2. Statistics and Significance Tests for Word Gap (Duration between the offset of the previous word and onset of the current word) for probable and improbable Words.

References

1. C. E. Shannon, *A mathematical theory of communication* (1948).

2. T. M. Cover, J. A. Thomas, *Elements of Information Theory* (John Wiley & Sons, 2012).

3. M. H. Goldstein, A. P. King, M. J. West, Social interaction shapes babbling: testing parallels between birdsong and speech. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 8030–8035 (2003).

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

5. J. Hale, “A Probabilistic Earley Parser as a Psycholinguistic Model” in *Second Meeting of the North AMerican Chapter of the Association for Computational Linguistics* (2001; https://aclanthology.org/N01-1021).

6. A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners. *OpenAI blog* **1**, 9 (2019).

7. T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, Others, Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).

8. A. Goldstein, Z. Zada, E. Buchnik, M. Schain, A. Price, B. Aubrey, S. A. Nastase, A. Feder, D. Emanuel, A. Cohen, A. Jansen, H. Gazula, G. Choe, A. Rao, C. Kim, C. Casto, L. Fanda, W. Doyle, D. Friedman, P. Dugan, L. Melloni, R. Reichart, S. Devore, A. Flinker, L. Hasenfratz, O. Levy, A. Hassidim, M. Brenner, Y. Matias, K. A. Norman, O. Devinsky, U. Hasson, Shared computational principles for language processing in humans and deep language models. *Nat. Neurosci.* **25**, 369–380 (2022).

9. A. Goldstein, A. Grinstein-Dabush, M. Schain, H. Wang, Z. Hong, B. Aubrey, M. Schain, S. A. Nastase, Z. Zada, E. Ham, A. Feder, H. Gazula, E. Buchnik, W. Doyle, S. Devore, P. Dugan, R. Reichart, D. Friedman, M. Brenner, A. Hassidim, O. Devinsky, A. Flinker, U. Hasson, Alignment of brain embeddings and artificial contextual embeddings in natural language points to common geometric patterns. *Nat. Commun.* **15**, 2768 (2024).

10. M. Heilbron, K. Armeni, J.-M. Schoffelen, P. Hagoort, F. P. de Lange, A hierarchy of linguistic predictions during natural language comprehension. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2201968119 (2022).

11. J. Raugel, Decoding of hierarchical inference in the human brain during speech processing with large language models (2024). https://2024.ccneuro.org/pdf/483_Paper_authored_CCN_abstract_final.pdf.

12. C. M. Brown, P. Hagoort, The neurocognition of language. *J. Psychophysiol.* **15**, 48–48 (2000).

13. G. K. Anumanchipalli, J. Chartier, E. F. Chang, Speech synthesis from neural decoding of spoken sentences. *Nature* **568**, 493–498 (2019).

14. T. Proix, J. Delgado Saa, A. Christen, S. Martin, B. N. Pasley, R. T. Knight, X. Tian, D. Poeppel, W. K. Doyle, O. Devinsky, L. H. Arnal, P. Mégevand, A.-L. Giraud, Imagined speech can be decoded from low- and cross-frequency intracranial EEG features. *Nat.*

Commun. **13**, 48 (2022).

15. A. Goldstein, H. Wang, L. Niekerken, Z. Zada, B. Aubrey, T. Sheffer, S. A. Nastase, H. Gazula, M. Schain, A. Singh, A. Rao, G. Choe, C. Kim, W. Doyle, D. Friedman, S. Devore, P. Dugan, A. Hassidim, M. Brenner, Y. Matias, O. Devinsky, A. Flinker, U. Hasson, Deep speech-to-text models capture the neural basis of spontaneous speech in everyday conversations, *bioRxiv* (2023)p. 2023.06.26.546557.

16. M. Kutas, K. D. Federmeier, Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annu. Rev. Psychol.* **62**, 621–647 (2011).

17. J. Polich, Updating P300: an integrative theory of P3a and P3b. *Clin. Neurophysiol.* **118**, 2128–2148 (2007).

18. P. Norvig, Natural language corpus data. *Beautiful data*, 219–242 (2009).

19. S. Garrod, M. J. Pickering, Why is conversation so easy? *Trends Cogn. Sci.* **8**, 8–11 (2004).

20. T. J. M. Sanders, W. P. M. Spooren, L. G. M. Noordman, Toward a taxonomy of coherence relations. *Discourse Process.* **15**, 1–35 (1992).

21. C. Clifton Jr, L. Frazier, Should given information come before new? Yes and no. *Mem. Cognit.* **32**, 886–895 (2004).

22. R. Giora, *On Our Mind: Salience, Context, and Figurative Language* (Oxford University Press on Demand, 2003).

23. N. C. Ellis, R. Simpson-vlach, C. Maynard, Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Q.* **42**, 375–396 (2008).

24. I. Kecskes, *The Socio-Cognitive Approach to Communication and Pragmatics* (Springer Nature, 2024).

25. Z. Zada, A. Goldstein, S. Michelmann, E. Simony, A. Price, L. Hasenfratz, E. Barham, A. Zadbood, W. Doyle, D. Friedman, P. Dugan, L. Melloni, S. Devore, A. Flinker, O. Devinsky, S. A. Nastase, U. Hasson, A shared model-based linguistic space for transmitting our thoughts from brain to brain in natural conversations. *Neuron*, doi: 10.1016/j.neuron.2024.06.025 (2024).

26. Z. M. Griffin, K. Bock, Constraint, Word Frequency, and the Relationship between Lexical Processing Levels in Spoken Word Production. *J. Mem. Lang.* **38**, 313–338 (1998).

27. R. C. Oldfield, A. Wingfield, Response latencies in naming objects. *Q. J. Exp. Psychol.* **17**, 273–281 (1965).

28. G. W. Beattie, B. L. Butterworth, Contextual Probability and Word Frequency as Determinants of Pauses and Errors in Spontaneous Speech. *Lang. Speech* **22**, 201–211 (1979).

29. R. Tikochinski, A. Goldstein, Y. Yeshurun, U. Hasson, R. Reichart, Perspective changes in human listeners are aligned with the contextual transformation of the word embedding

space. *Cereb. Cortex* **33**, 7830–7842 (2023).

30. C. J. Honey, T. Thesen, T. H. Donner, L. J. Silbert, C. E. Carlson, O. Devinsky, W. K. Doyle, N. Rubin, D. J. Heeger, U. Hasson, Slow cortical dynamics and the accumulation of information over long timescales. *Neuron* **76**, 423–434 (2012).

31. J. R. Manning, J. Jacobs, I. Fried, M. J. Kahana, Broadband shifts in local field potential power spectra are correlated with single-neuron spiking in humans. *J. Neurosci.* **29**, 13613–13620 (2009).

32. T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, HuggingFace's Transformers: State-of-the-art Natural Language Processing, *arXiv [cs.CL]* (2019). http://arxiv.org/abs/1910.03771.