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

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 Casto[4,5], Werner Doyle[6], Daniel Friedman[6], Sasha Devore[6], Patricia Dugan[6], Avinatan Hassidim[2], 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

The neural basis of spontaneous speech production, in which speakers efficiently and effortlessly generate utterances on the fly to express their thoughts, is among the least understood aspects of human cognition. This study utilizes information theory, contemporary Large Language Models (LLMs), and approximately 100 hours of high-quality spatiotemporal ECoG recordings of speakers engaged in spontaneous conversations to explore how the speaker's brain conveys information during everyday interactions. Information theory defines information as the reduction of uncertainty (Shannon entropy). It lays the theoretical foundations for why listeners actively predict upcoming words (information-seeking) before a word is spoken while enhancing the processing of unexpected, information-rich words after they are perceived. But what happens when speakers generate (information-making) these improbable, information-rich words in the first place? We analyzed continuous electrocorticography (ECoG) recordings collected during hours of real-life, 24/7 conversations to address this question. Using LLMs (Llama-2 and GPT-2), we estimated the probability of each word based on its context, categorizing them as either improbable, information-rich words or predictable, information-thin words. We then extracted word-based non-contextual embeddings from these models and employed neural encoding techniques to examine brain activity during speech production and comprehension. Our findings reveal a striking contrast in how the brain handles improbable, information-rich words while speaking versus listening. During speech comprehension, we identified two distinct neural phases: one preceding word onset, associated with predictive (information-seeking) processing, and another following word onset, linked to enhanced information processing of unexpected words. Conversely, in the speaker's brain, we found, for the first time, enhanced pre-word-onset encoding for improbable, information-rich words 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. Since information-rich words are statistically unpredictable, this suggests the speaker's brain aims to produce linguistic output that defies listeners' expectations. However, we also point out that predictability alone is insufficient to generate meaningful words, highlighting a gap in information theory and LLMs that neglects how speakers intentionally choose information-rich words to convey novel meanings.

Introduction

The neural basis of spontaneous speech production is one of the least studied and understood aspects of human cognition. In everyday conversations, people quickly generate sequences of utterances, often without being aware of how they select the words they use to express their thoughts. In contrast, most research on speech production has focused on highly controlled and predetermined sequences, frequently neglecting the complexities involved in spontaneous speech (1–3). This study utilizes a unique dataset of approximately 100 hours of high-quality spatiotemporal ECoG recordings, where speakers engage in spontaneous conversations, to investigate how they convey information to listeners during daily interactions.

In information theory, information is formally defined in terms of reduction in uncertainty (represented by Shannon entropy, (4–6)) in a received message. 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 (7, 8). Until the recent advancements in language modeling, estimating the amount of information each word provides during natural conversations was nearly impossible. The ability to train a large language model to predict the next word based on prior context, utilizing all available text from the internet, provided a new tool to evaluate the amount of information conveyed during any natural conversation. Specifically, Large Language Models (LLMs) give a probability score for the likelihood of saying each word in the lexicon at any given moment in a conversation, given all prior words (context) and prior knowledge accumulated in the network. Previous research shows that the level of certainty of LLMs regarding the next word aligns with listeners' certainty levels as they process natural language (9, 10)

Although information theory predicts how neural responses should adapt during speech comprehension (information-seeking), we still lack insights into how it shapes information generation during speech production (information-making). During speech comprehension before word-onset, information theory demonstrates that listeners actively predict the next word before it is spoken (10, 11). After word onset, information theory indicates that the brain estimates the surprise level (prediction error) and processes the surprising (informative) words (10, 12, 13). However, how information theory and Shannon entropy influence spontaneous speech production is unclear. On the one hand, entropy may have a minimal impact on speech production, as speakers often have knowledge and minimal uncertainty about the words they intend to say. On the other hand, since information is key in any conversation, the speaker's brain may engage in more effortful and deliberate processes when producing less probable, information-rich words compared to more predictable, less informative ones.

To address this question, we asked how information modifies the neural responses during speech production and comprehension. 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 (14). 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. Moreover, 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 and after word onset in the same individuals during speech production and comprehension. This is as opposed to prior research, which, due to the limited SNR for high-gamma power (correlated with average fringing rate - cite) of EEG and MEG methods, has primarily focused on assessing the post-word-onset surprise effect, such as the P300 and N400 markers (15, 16).

To quantify the amount of information conveyed by each word in hundreds of recorded conversations, we used two language models (LLMs), specifically Llama-2 and GPT-2. These models assigned a probability (certainty level, ranging from zero to one) to say each word based on the context of all the preceding words in each conversation. We categorized all spoken words based on their probability scores. We separated them into two groups: predictive words with low entropy (top 30% probable, as predicted by LLMs) and surprising words with high entropy (the bottom 30% improbable, which the LLMs did not predict well). 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 (information-thin) and improbable (information-rich) words in natural, real-life conversations.

Our key findings suggest opposing, perhaps complementary, functionality of the language areas before word onset when engaged with listening compared to speaking. In listeners' IFG, we reproduced our previous finding of enhanced pre-word-onset encoding in speech comprehension (10, 11) for probable versus improbable words. 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 slowed down their 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 indicate that although information-rich words are relatively rare in conversation, occurring less than 30% of the time, they enhance the neural responses in the speaker's brain across many language areas during speech production. Since information-rich words are statistically unpredictable, it implies that the information-making processes in the speaker's brain strive to generate linguistic output that challenges the listener's expectations. This challenges the idea that probabilistic speech, as implemented in many LLMs, is sufficient for capturing the complexities of human language generation.

Results

Our 24/7 conversation data consists of half a million words recorded during 100 hours of spontaneous conversations between four ECoG patients and their surroundings in the hospital room. The conversations cover various real-life topics, including discussions between the patients and medical staff and personal conversations about family, friendships, sports, and politics. We utilized this rare dataset to evaluate how speakers communicate information to listeners during natural conversations.

To measure the amount of information communicated through words in our conversations, we utilized LLMs (Llama-2 and GPT-2) to assess the predictability of each word based on each conversational context. Our analysis of our 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 harder to predict. 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 (Supp. Fig. 1).

Using two metrics derived from LLMs, we categorized word instances into information-rich improbable words (bottom 30%) and information-thin probable words (top 30%). The first metric was based on the prediction accuracy of the LLMs, and the second was based on the LLMs' confidence levels (see Supp. Table 1). We trained encoding models to predict the neural responses to improbable (information-rich) and probable (information-thin) words in each conversation. Each word was represented by a non-contextual word embedding extracted from the LLMs' non-contextual embedding layer. We trained independent encoding models for each electrode during speech comprehension and speech production at time lags ranging from –2 s to +2 s relative to the onset of words (lag 0).

During speech comprehension (listening), we identified two neural phases. The first phase occurred before word onset, during which listeners actively sought information by predicting the next word in the conversation. During this phase, we observed enhanced encoding for upcoming probable words compared to improbable words 100 to 500 milliseconds before word onset (see Fig. 1A). Interestingly, the predictive signals were primarily localized to the inferior frontal gyrus (IFG, also known as Broca's area; see the red electrodes in the brain map in Figure 1A). This finding replicates our recent discovery of enhanced encoding for probable words before word onset during continuous listening to a podcast (10), now using new data from spontaneous conversations. This result supports the claim that during speech comprehension (information-seeking), the brain makes predictions based on language statistics (as coded in LLMs).

The second phase occurred after word onset, during which listeners actively engaged in additional processing for improbable (information-rich) words. After word onset, we observed a shift in the response patterns in the listener's brain, where we detected enhanced encoding around 100 to 500 milliseconds after word onset for the improbable (surprising) words (see Fig. 1B). In contrast to the localized predictive response before word onset, the enhanced processing for improbable (surprising) words after word onset was widespread (Out of 119 electrodes identified as significant for speech

processing (see Materials and Methods), 52 exhibited a significant preference for improbable words, while 13 showed a preference for probable words. The proportions were found to be significantly different ($p < .001$). see blue electrodes in the brain map in Figure 1B). Together, these results provide direct evidence that the listener's brain is tuned to detect (before word onset) and process (after word onset) information-rich words during comprehension.

We observed a complementary neural pattern associated with information generation during speech production in the speaker's brain before word onset. In contrast to the pre-onset enhanced encoding for probable words in the listener's brain, we observed enhanced encoding for improbable (surprising) words over probable words in language areas (Fig. 1C). The enhanced encoding for improbable words in the speaker's brain was significant and widespread across multiple language regions. Out of 123 significant electrodes for speech processing, 110 exhibited a marked preference for unlikely words, while only 4 showed a preference for likely words. The proportions were found to be significantly different ($p < .001$, see blue electrodes in the brain map in Figure 1C, see also additional ROIs in Supp. Fig. 2). The enhanced encoding for improbable, information-dense words was still apparent in the speaker's brain for 100 to 500 ms (green bar) after the words are spoken (Figure 1D). The results indicate that before word onset, there are stronger, more easily decoded neural signals for information-rich improbable words during speech production, suggesting that increased processing is involved in their formation.

The contrast between enhanced encoding for improbable words in the speaker's brain (information-making) and enhanced encoding for probable words (information-seeking) 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 with instances 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' induced probability 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).

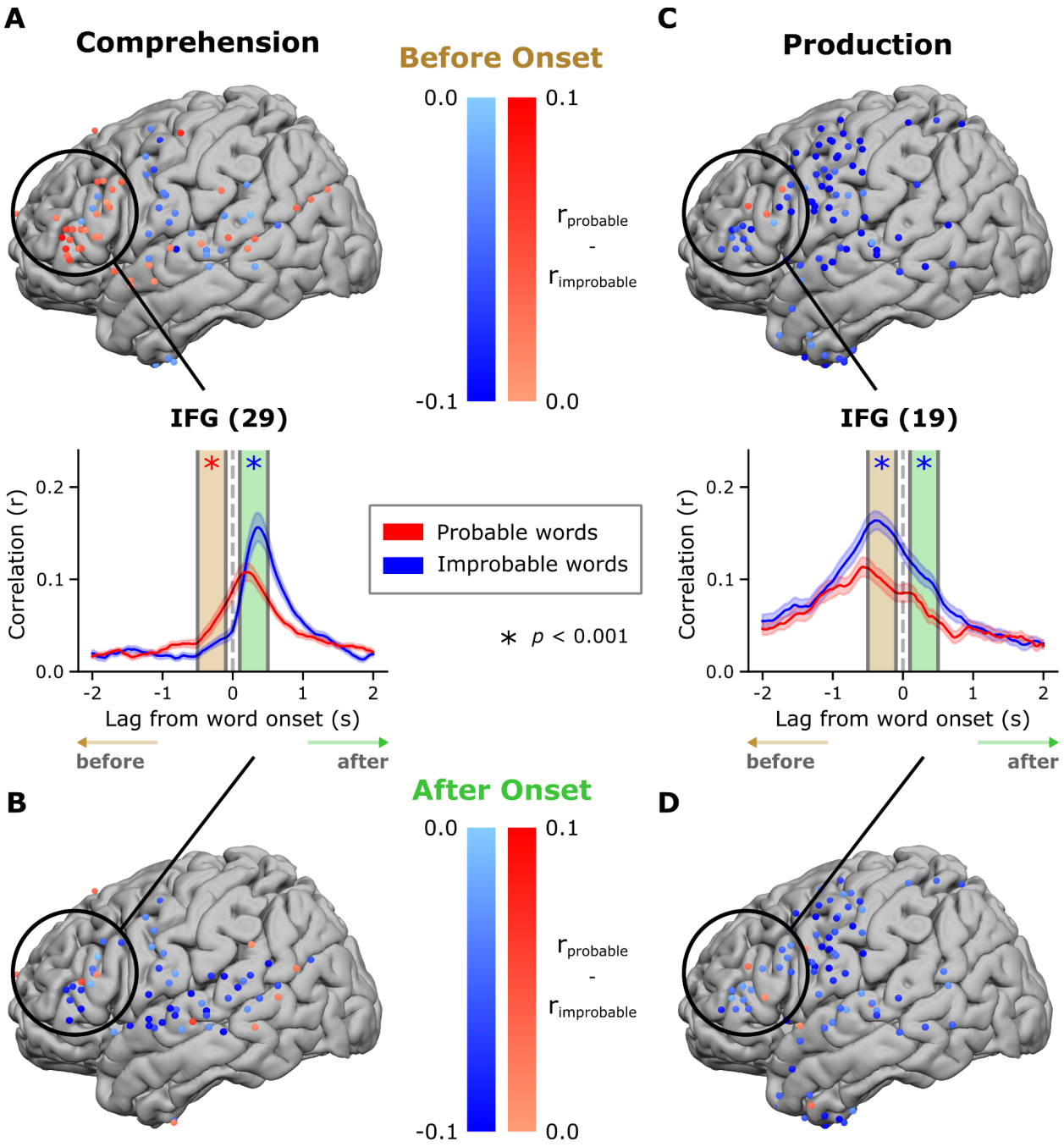


Figure 1. Electrodwise and IFG-localized encoding for Probable and Improbable Words. A. Before-word onset information-seeking processed during speech comprehension (listening). Using LLM (Llama-2) embeddings, we observed enhanced encoding in the listeners' brains for probable (predictable) words (red) compared to improbable (surprising) words (blue) around 100 to 500 ms (brown bar) before the words were spoken. The enhanced encoding for probable words was predominantly localized to the inferior frontal gyrus (IFG, also known as Broca's area). The enhanced encoding for probable (predictable) words suggests that the listener's brain anticipates the subsequent utterances before they are articulated. **B.** After-word onset information-processing during speech comprehension (listening). We observed a widespread enhanced encoding in the listeners' brains for improbable (surprising) words (blue) compared to probable (predictable) words (red) around 100 to 500 ms (green bar) after the words were spoken. The enhanced encoding for improbable (information) words suggests that the listener's brain is engaging in additional processing of the

improbable, information-rich utterances after they are perceived. **C.** Before-word onset information-making processes during speech production (speaking). We observed enhanced encoding in the speaker's brain for improbable (surprising) words (blue) compared to probable (predictable) words (red) around 100 to 500 ms (brown bar) before the words were spoken. The enhanced encoding for probable words was predominantly localized to the inferior frontal gyrus (IFG, also known as Broca's area). The enhanced encoding for improbable words suggests that the speaker's brain engages in additional processing while producing information-rich utterances. **D.** The enhanced encoding for improbable, information-rich words remained high and significant in the speaker's brain even 100 to 500 ms (green bar) after the words were spoken. 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).

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, information-rich 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 (17) 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).

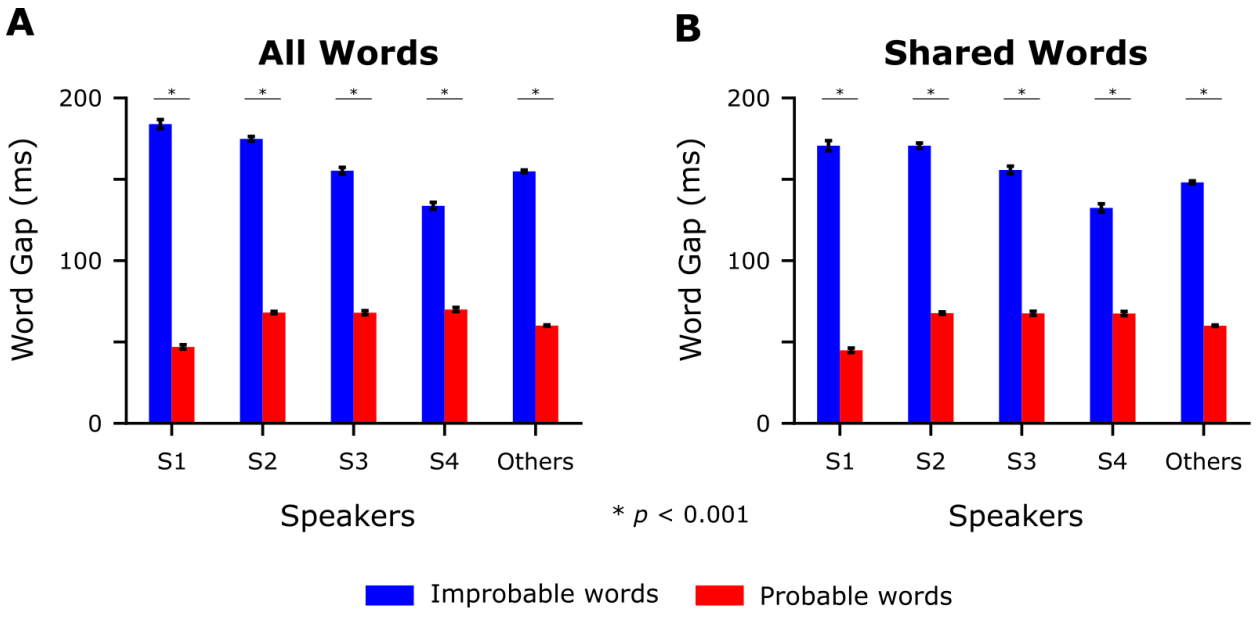


Figure 2. Behavior Temporal Gap between the Offset of the Previous Word and Onset of the Current Word. A. It takes about 100 - 150 ms longer for each speaker (S1- S4) to start articulating improbable words. This was also evident when examining the data of other speakers in the room, for whom brain responses 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

In everyday conversations, listeners often aim to acquire new information they do not possess and, thus, can not easily predict. After all, if listeners could fully predict the speaker's utterances, there was little point in engaging in the conversation. In line with the information theory perspective, our analysis of 100 hours of recordings shows that a subset of approximately 20-30% of all spoken words in our recorded conversations were improbable and information-rich. However, Little is known about the underlying neural processes in the speaker's brain that generate these information-rich words. By measuring the neural activity in the speaker's brain during hundreds of real-life conversations, we have discovered evidence for a novel information-making process in the speaker's brain before articulating words, complementing the information-seeking processes in listeners' brains.

Behaviorally, our analysis found that speech production slows by about 100 to 150 ms before articulating improbable words (Fig. 2). This is consistent with previous research that has indicated that speakers take longer to begin articulating rare (infrequent) words (18, 19). This process is often attributed to difficulty retrieving infrequent words from memory or planning an articulatory motor sequence (20). Our findings suggest that the effect is context-specific rather than word frequency-specific, as the pause in articulation can be longer for the exact words spoken in unpredictable versus predictable contexts. Besides the increased gap before articulating improbable words, it was established that the duration of articulating each word also slows down (21). The results may seem counterintuitive, given that the speaker's brain already knows the words they are about to produce, irrespective of how surprising they are from the listener's perspective. Thus, the slowdown in speech rate may provide a behavioral marker for additional cognitive processes in the speaker's brain while producing information-rich (improbable) words.

Our analysis of encoding in the speaker's brain shows additional neural processes involved in producing information-rich, improbable words compared to probable ones (Fig. 1C). The enhanced encoding for improbable words before word onset in the speaker's brain contrasts with the enhanced encoding for probable words that occur before word onset in the listener's brain (Fig. 1A). Furthermore, while predictive information-seeking processes in the listener's brain were localized to the IFG (Fig. 1A), the information-making processes in the speaker's brain were widespread across various language areas, including the IFG, STG, angular gyrus, and precentral motor cortex (Fig. 1C and Supp. Fig. 2).

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

Speech comprehension and speech production offer complementary perspectives on the relationship between information theory and LLMs to natural language processing. In speech comprehension, the primary focus is on how listeners process linguistic information from real-world linguistic input. Recent evidence suggests that, like LLMs, humans rely on a next-word predictive coding framework to compress linguistic information into an embedding space (10, 11, 14, 22). Thus, in alignment with

information theory, listeners constantly seek to detect information-rich words to adjust their linguistic model based on their prediction errors.

Speech production offers a complementary viewpoint that explores how speakers choose the right words to express their thoughts and ideas. This process can be formalized as the policy for selecting the next word (or sequence of words) from the statistically learned language model given each conversational context. It was shown that forcing LLMs to choose the most predictable (top-1) word in any conversational context leads to repetitive, incoherent, and uninformative speech (23). Thus, like humans, LLMs must generate some information-rich and surprising words during speech production. However, they should not solely aim to maximize surprise, as this approach may lead to the selection of highly improbable (essentially random) words. Many language models use temperature-controlled sampling methods and other strategies to sound more human-like. These techniques allow them to select words based on probabilities from a distribution without further analysis to distinguish between information-rich and information-thin words (24).

Our discovery of additional neural processes in the speaker's brain for generating information-rich (surprising) words exposed a gap in information theory. While information theory focuses on the transfer of information between a sender and receiver over a noisy channel, it does not adequately address how humans generate new (surprising) ideas. While entropy can indicate how much new information a listener is receiving, it doesn't offer much insight into how a speaker thoughtfully chooses words to convey new meanings. The lack of dedicated information-making processes during speech production may explain why LLMs deteriorate when trained with text generated by other LLMs rather than humans (23). In such cases, the generated probabilistic text becomes more predictable and less informative over time, leading to a rapid deterioration of the language model. In other words, the probabilistic policy for speech production used by LLMs struggles to capture the communicative intent that guides speakers in selecting the key informative words that reflect their thoughts and ideas. After all, listeners are not likely to randomly choose the key information-rich words they wish to convey to their audience. In agreement with this intuition, our findings suggest an additional information-making mechanism in the human brain that requires extra neural resources and time to select and produce key information-rich words during natural conversations.

This study has several limitations. First, the nature of the neural policy in the speaker's brain associated with choosing improbable yet informative words is not yet defined. While we observe the improved encoding of improbable words in the speaker's brain, we know very little about the underlying policies 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 linked to the human capacity to think and innovate. Furthermore, while information-seeking and entropy are fundamental for understanding speech comprehension, entropy gives us only a narrow window into speech production.

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 (10). 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 an overestimation of the level of surprise for the bottom 30% of the words. Given that some of these words may be less surprising for our listeners, this should reduce (not enhance) the gap between surprising and unsurprising words observed in our study (Fig. 1).

To conclude, the novel 24/7 ECoG recordings of natural conversation provide a new window to novel information-making processes in the speakers' brains, complementing the proposed 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 underutilized in most LLMs, which rely solely on probabilistic speech to generate conversations. Indeed, recent research in deep learning seeks to develop better policies for sampling words during speech production by relying on the internal chain of thought processes to generate a more thoughtful response in context. Such information-making processes may be the key to understanding how we 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 (14).

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 (25). Neural signals were bandpassed (75–200 Hz), and power envelopes were computed as proxies for local neural firing rates (26). 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 (14).

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 (27). We first converted the words from the raw transcript (including punctuation and capitalization) to whole 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 (10).

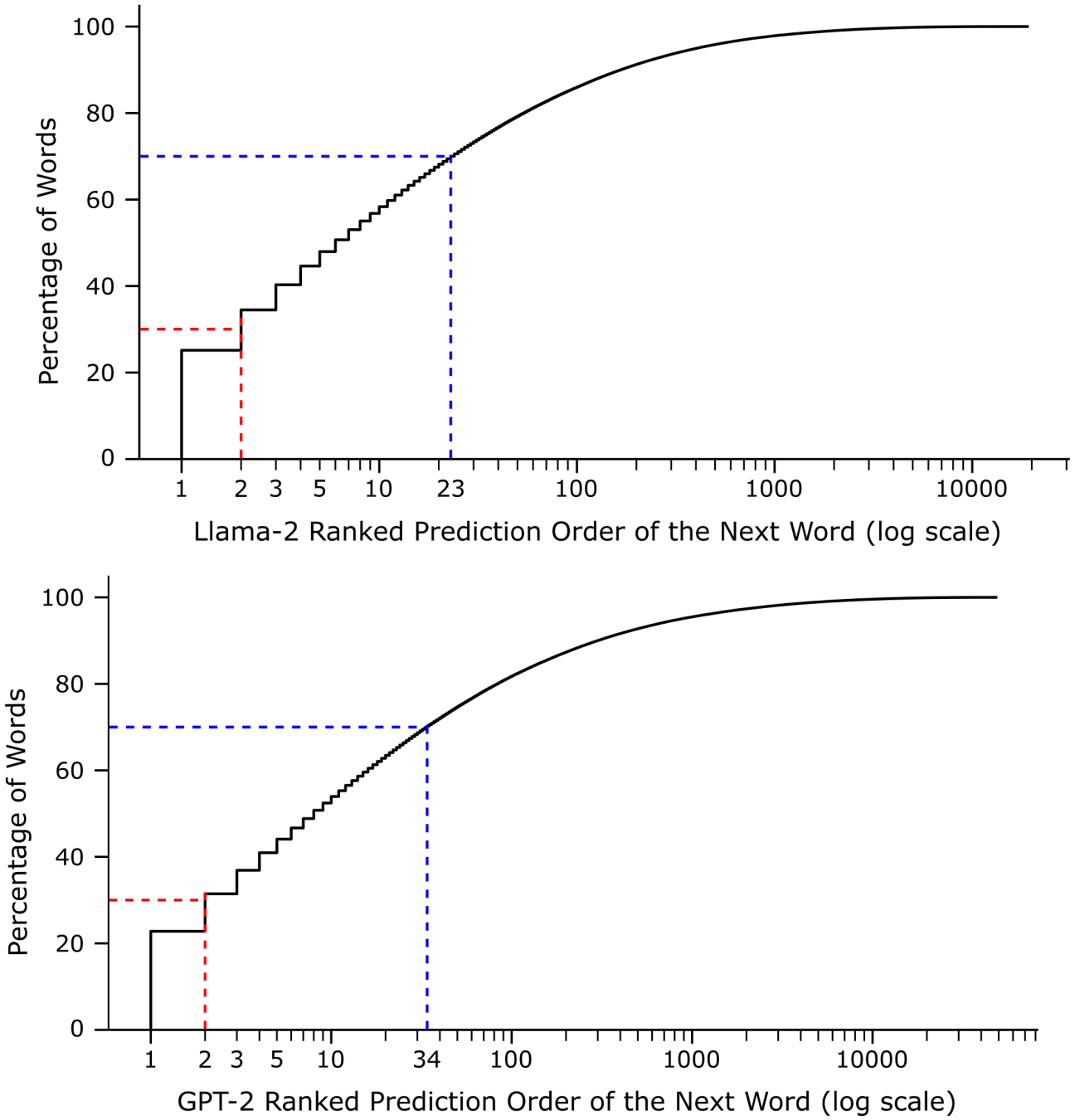
Electrode selection

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 (10).

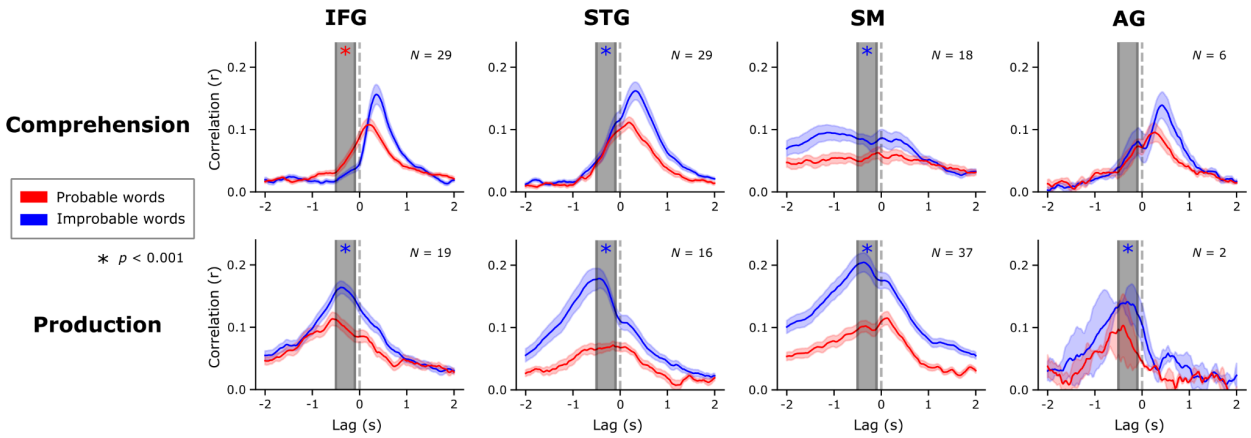
Significance test for encoding difference at the ROI level

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.

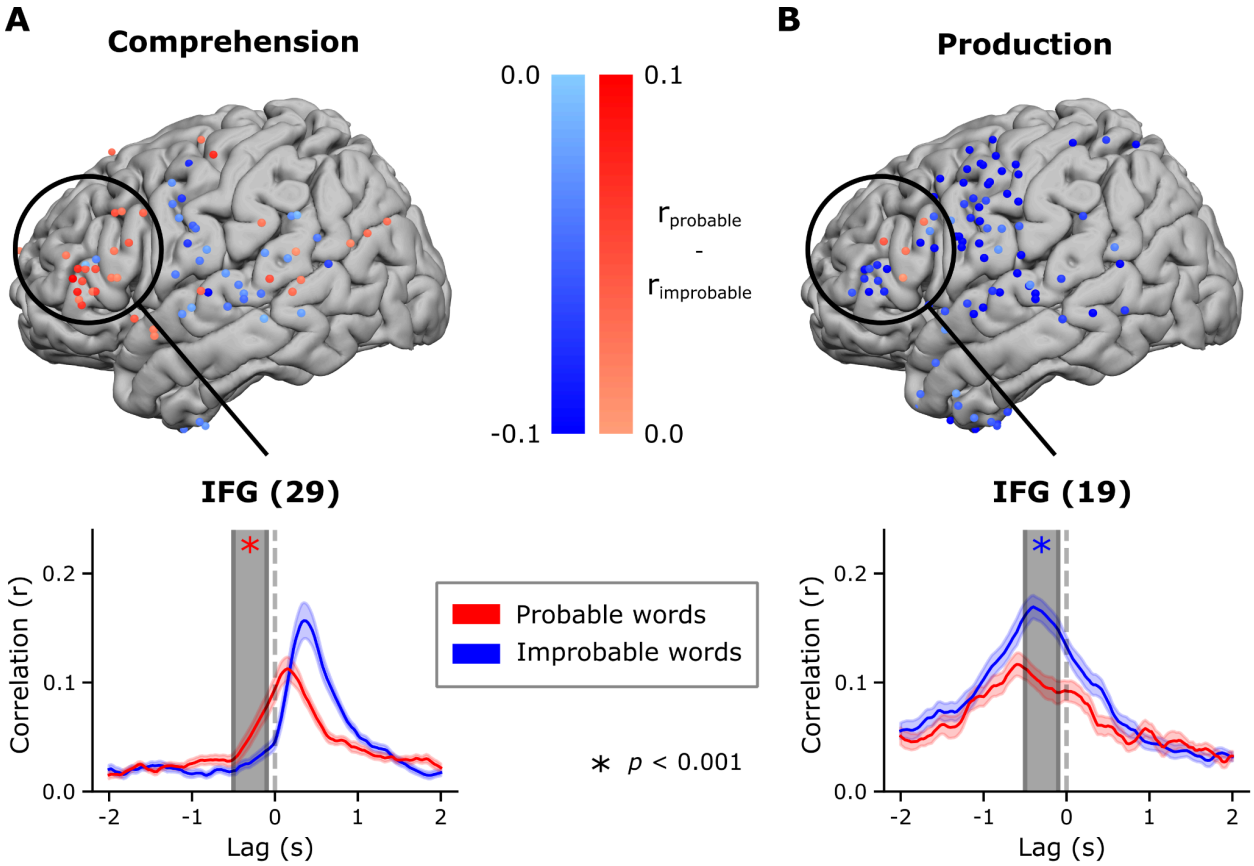
Supplementary Figures



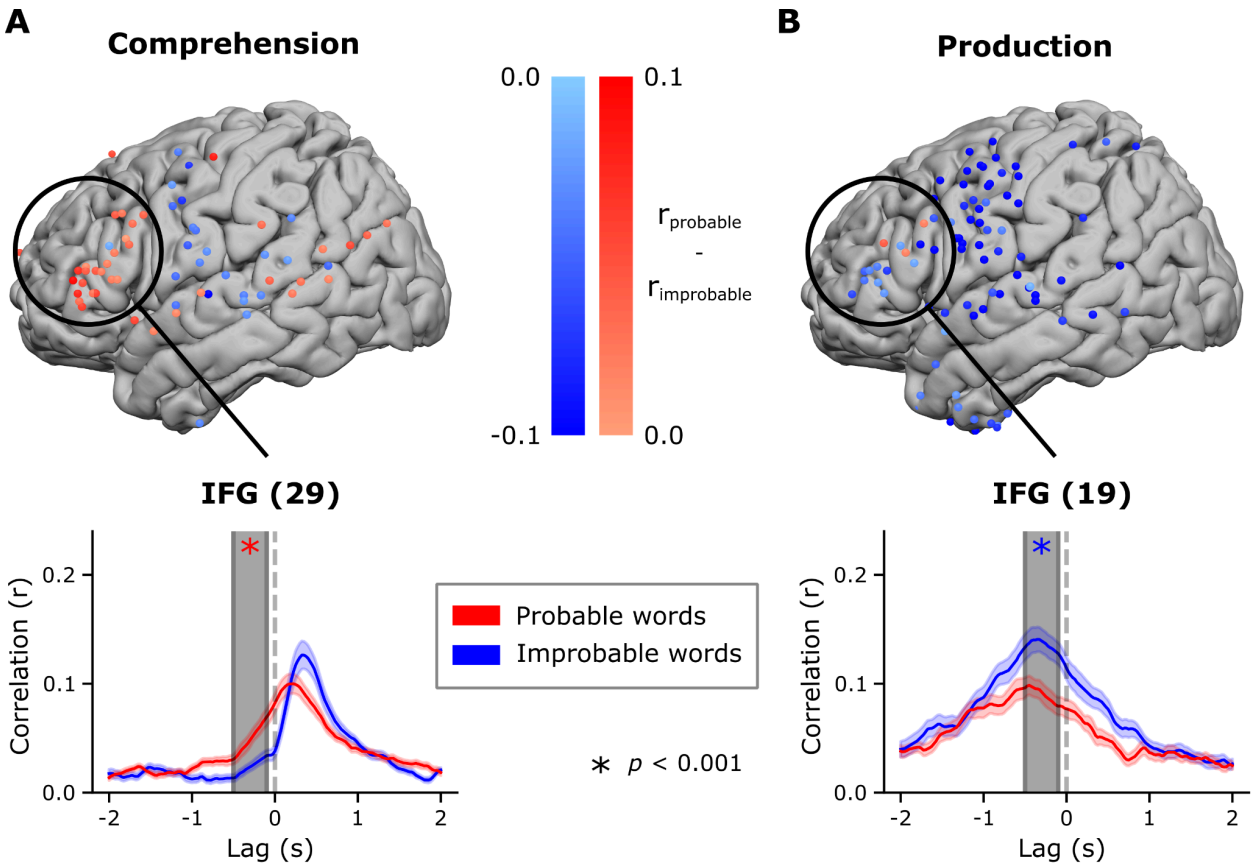
Supplementary Figure 1. Accumulated ranked-order predictions for upcoming words as predicted by Llama-2 and GPT-2. We extracted each next word's ranked probability according to Llama-2 (up) and GPT-2 (bottom) context-based predictions. The rank order is represented on a logarithmic scale. LLMs successfully 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%.



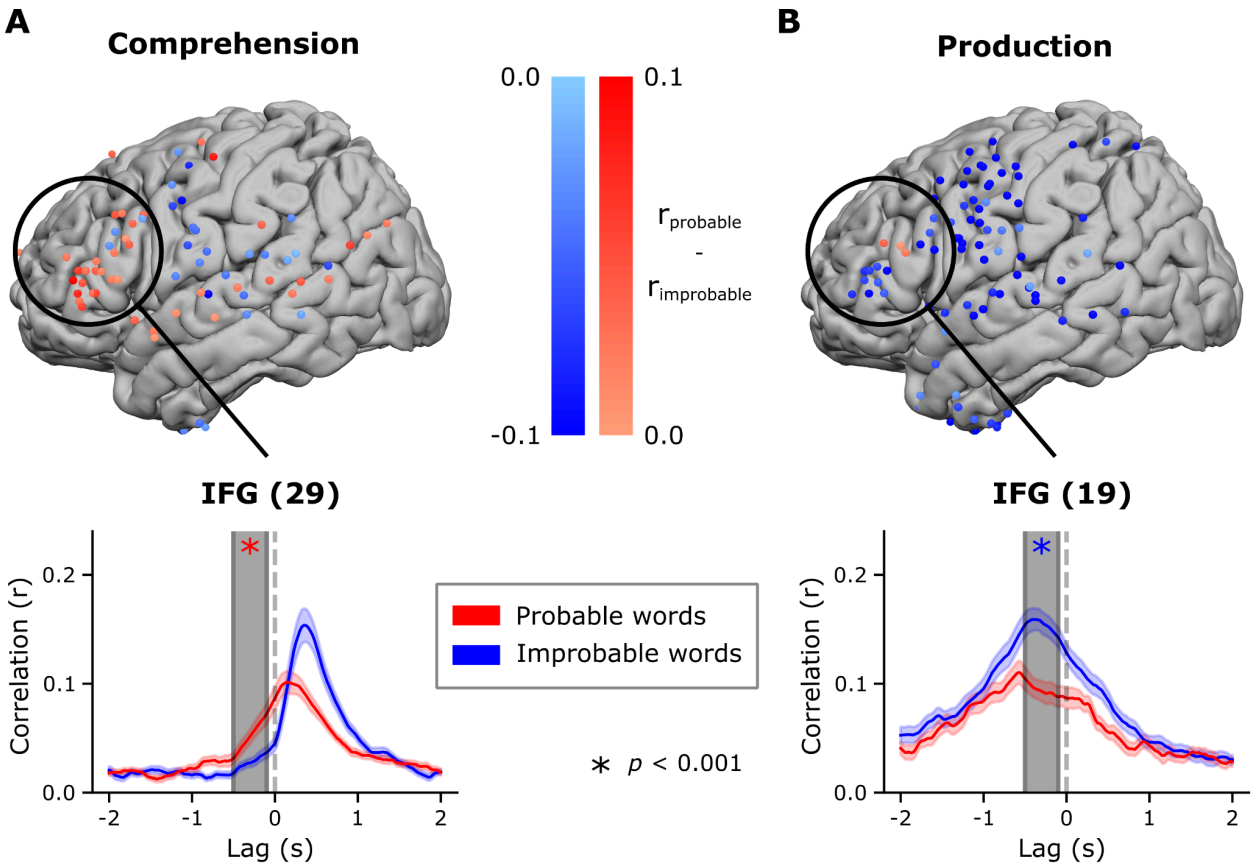
Supplementary Figure 2. Encoding Results for Probable and Improbable Words in Different ROIs. The listener's brain showed enhanced pre-word-onset encoding of probable words in the IFG, while the speaker's brain exhibited widespread enhanced pre-word-onset encoding of improbable words across several language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



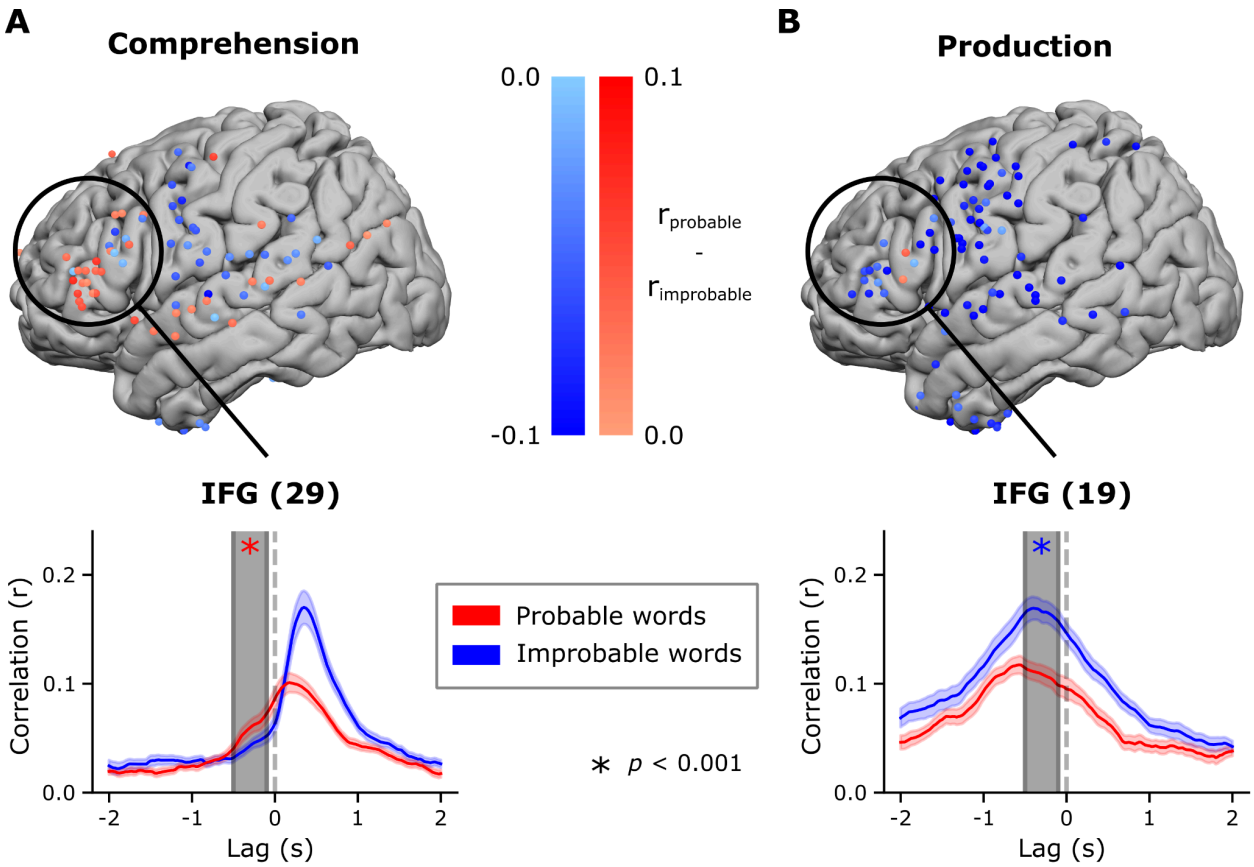
Supplementary Figure 3. Using content words to encode probable and improbable words. Similar results to those in Fig. 1 were achieved while restricting the encoding analyses to content words (i.e., nouns, verbs, adjectives, and adverbs, N = 306,681). This demonstrates that highly predictable function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



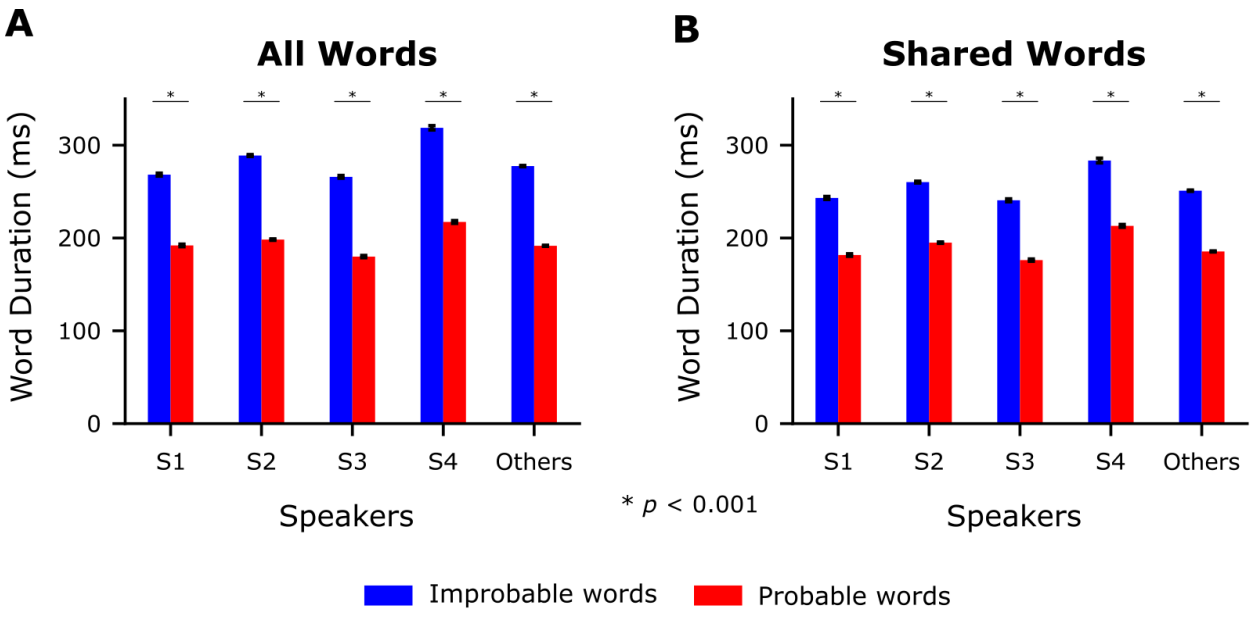
Supplementary Figure 4: Utilizing a shared set of words to encode probable and improbable words. Similar results to those in Figure 1 and Supplementary Figure 3 were achieved using a shared set of words, which were predictable in one context and unpredictable in another. This demonstrates that the observed 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 the same set of words that 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. M. Brown, P. Hagoort, The neurocognition of language. *J. Psychophysiol.* **15**, 48–48 (2000).

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

3. T. Proix, *et al.*, Imagined speech can be decoded from low- and cross-frequency intracranial EEG features. *Nat. Commun.* **13**, 48 (2022).

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

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

6. 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).

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

8. 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).

9. D. Sivan, M. Tsodyks, Information theory of meaningful communication. *arXiv [cs.CL]* (2024).

10. A. Goldstein, *et al.*, Shared computational principles for language processing in humans and deep language models. *Nat. Neurosci.* **25**, 369–380 (2022).

11. A. Goldstein, *et al.*, Alignment of brain embeddings and artificial contextual embeddings in natural language points to common geometric patterns. *Nat. Commun.* **15**, 2768 (2024).

12. 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).

13. J. Raugel, Decoding of hierarchical inference in the human brain during speech processing with large language models. (2024). Available at: https://2024.ccneuro.org/pdf/483_Paper_authored_CCN_abstract_final.pdf [Accessed 30 July 2024].

14. A. Goldstein, *et al.*, Deep speech-to-text models capture the neural basis of spontaneous speech in everyday conversations. *bioRxiv* 2023.06.26.546557 (2023).

15. 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).

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

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

18. 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).

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

20. P. Mousikou, K. Rastle, Lexical frequency effects on articulation: a comparison of picture naming and reading aloud. *Front Psychol* **6**, 1571 (2015).

21. 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).

22. Z. Zada, *et al.*, A shared model-based linguistic space for transmitting our thoughts from brain to brain in natural conversations. *Neuron* **112**, 3211–3222.e5 (2024).

23. I. Shumailov, *et al.*, AI models collapse when trained on recursively generated data. *Nature* **631**, 755–759 (2024).

24. A. Holtzman, J. Buys, L. Du, M. Forbes, Y. Choi, The curious case of neural text degeneration. *arXiv [cs.CL]* (2019).

25. C. J. Honey, *et al.*, Slow cortical dynamics and the accumulation of information over long timescales. *Neuron* **76**, 423–434 (2012).

26. 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).

27. T. Wolf, *et al.*, HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv [cs.CL]* (2019).