# DIFFUSION-BASED DYNAMICS AS A COGNITIVE MODEL OF HUMAN SPEECH PRODUCTION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Human language production requires transforming abstract communicative intent into fluent speech, yet the algorithmic nature of this transformation remains less understood. Most studies aligning large language models (LLMs) with brain activity have focused on autoregressive LLMs (aLLMs), which generate text left-to-right by committing to the next token. While effective at predicting neural and behavioral signatures of comprehension, this paradigm assumes incremental generation. In contrast, diffusion LLMs (dLLMs) construct sentences by iteratively denoising global representations. Despite their distinct generative dynamics, dLLMs now rival aLLMs on standard NLP benchmarks, prompting the question of whether the brain likewise engages in global, iterative refinement—especially during pre-articulatory planning when sentence structure remains flexible. To test this hypothesis, we correlated intermediate denoising steps of a dLLM with electrocorticography (ECoG) activity during naturalistic speech production. dLLM representations explained significant neural variance from pre- to post-production, with especially strong encoding in middle/inferior temporal and motor-related regions. These results support iterative refinement as a plausible neural mechanism of human speech planning.

## 1 INTRODUCTION

Human language processing—from comprehension to internal formulation to overt production—is a window into the mind's generative machinery. Most large language models (LLMs) used in model-brain alignment studies operate via a left-to-right, next-token prediction paradigm (Caucheteux et al., 2023; Gao et al., 2025; Goldstein et al., 2022; 2025; Schrimpf et al., 2021; Toneva & Wehbe, 2019; Antonello et al., 2024; Jain & Huth, 2018). These autoregressive architectures have proven surprisingly effective at capturing aspects of human brain activity during naturalistic language tasks, especially when scaled up (Gao et al., 2025; Antonello et al., 2024; Hong et al., 2024). However, they instantiate one specific algorithmic hypothesis about how linguistic output is constructed: sequential conditional commitment to the next word. Emerging diffusion LLMs (dLLMs), such as LLaDA (Nie et al., 2025) and Dream (Ye et al., 2025), propose a qualitatively different generative mechanism. Instead of predicting the next token, they begin from a noisy, underspecified representation and iteratively denoise toward a coherent sentence. Despite their contrasting dynamics, large-scale dLLMs now rival aLLMs such as LLaMA3 (Grattafiori et al., 2024) and Qwen2.5 (Yang et al., 2024) across a range of NLP benchmarks. This computational plurality challenges the assumption that next-word prediction is the sole viable substrate for language modeling.

The question arises: Could the brain's language production resemble an iterative refinement more than a left-to-right sequence? In this view, speakers hold a graded, probabilistic proto-utterance that is progressively refined before speech onset, with lexical, syntactic, and discourse constraints gradually resolving into a coherent plan for articulation. Here, we test the cognitive plausibility of this hypothesis by correlating intermediate embeddings of dLLMs (LLaDA and Dream) across denoising steps with electrocorticography (ECoG) activity during naturalistic speech production. To the best of our knowledge, no published study has yet reported direct comparisons of a dLLM's internal representations with neural recordings during language tasks – an important open research question. We find that (1) earlier diffusion steps preferentially generate high-frequency content words, diverging from the sequential patterns in autoregressive models; (2) dLLM embeddings evolve in parallel with neural dynamics in temporal and motor regions; and (3) earlier steps align with pre-

articulatory activity, while later steps align with motor execution and post-articulatory processing. Together, these results highlight diffusion models as a cognitively plausible framework for capturing the dynamics of human speech production.

## 2 Related Work

**Diffusion LLM.** Early diffusion models such as Diffusion-LM (Li et al., 2022) and D3PM (Austin et al., 2021) were relatively small, but they established the foundation for today's billion-parameter models such as LLaDA-8B (Nie et al., 2025) and Dream-7B (Ye et al., 2025). LLaDA adopts the core architecture of LLaMA3's transformer-based LLM. It uses the same tokenizer and Transformer layer stack as LLaMA3-8B, enabling direct performance comparisons. Crucially, LLaDA removes the causal (left-to-right) attention mask used in LLaMA models, allowing bidirectional self-attention over the sequence. A special masking token (<MASK>) is introduced into the vocabulary to represent "noisy" or corrupted tokens during diffusion-style generation. Dream's architecture is directly derived from Qwen2.5-7B and was initialized with Qwen2.5-7B's pretrained weights to bootstrap its knowledge. The key architectural modification for Dream was analogous to LLaDA's: switching from Qwen's causal masking to full bidirectional self-attention.

Both LLaDA and Dream depart from conventional autoregressive training in that they do not use next-word prediction on a prefix. Instead of producing one token at a time, they predict many tokens at once given a partially masked context. This means the loss is computed over multiple token positions simultaneously (all masked tokens) rather than only the next position. The training data for dLLMs also must include very high masking ratios (up to 100% masked) so that the model learns to generate whole sequences from nothing. Another important difference is the use of time-step conditioning in dLLMs: the model is aware of a "step" or mask level during training, which aLLMs do not require. This was implemented by adding an encoding of the mask fraction or diffusion step index into the model's input or hidden layers. The outcome is that dLLMs learn a sequence of denoising steps rather than a single-step distribution.

Both LLaDA and Dream have demonstrated that dLLMs can match or even surpass aLLMs on many language tasks. Industry efforts, such as Mercury (Labs et al., 2025) and Gemini Diffusion (Deepmind, 2024) report generation speeds of thousands of tokens per second using optimized parallel sampling, demonstrating that dLLMs are becoming practical alternatives rather than mere academic curiosities.

**Psycholinguistic models of speech production.** Psycholinguistic models of speech production (Levelt, 1989; 1999; Garrett, 1975; 1980; Dell, 1986; 1988) have long proposed that speaking proceeds through a hierarchy of stages. For example, Levelt's (1989; 1999) production model posits that speech unfolds through a serial process of conceptualizing a message, formulating it into linguistic form (lemmas, syntax, phonology), articulating it via motor commands, and monitoring output to ensure accuracy. Dell's (1986; 1988) interactive model proposes that speech production arises from spreading activation across semantic, lexical, and phonological levels, allowing information to cascade bidirectionally and explaining speech errors through simultaneous competition and interaction between representations.

Neuroimaging evidence complements these models by revealing multiple hierarchical levels in the brain's speech production network. Classical lesion studies established Broca's area as critical for articulation (Broca, 1861), and more recent work has shown that it orchestrates an orderly sequence of linguistic operations. Intracranial recordings in Broca's area, for instance, reveal a temporal cascade spanning word selection, grammatical encoding, and phonological encoding (Sahin et al., 2009). Similarly, motor cortex recordings have demonstrated structured maps and temporal codes for phonemes, reflecting the fine-grained organization of articulatory gestures (Bouchard et al., 2013).

**Model-brain alignment during language use.** In recent years, numerous studies have reported parallels between LLMs and human brain activity during language processing (Antonello et al., 2024; Caucheteux et al., 2023; Gao et al., 2025; Hong et al., 2024; Jain & Huth, 2018; Goldstein et al., 2022; 2025; Schrimpf et al., 2021; Toneva & Wehbe, 2019). For instance, GPT-2's word probabilities explained unique variance in ECoG responses in language areas, suggesting that both the brain and LLMs rely on predictive representations (Goldstein et al., 2022). Such findings support the idea

that the simple objective of next-word prediction captures important aspects of the brain's computation. More recent LLMs such as LLaMA (Touvron et al., 2023) and OPT (Zhang et al., 2022) have been shown to align more closely with brain activity during language processing, exhibiting a scaling law whereby larger models yield improved brain predictivity (Antonello et al., 2024; Hong et al., 2024; Gao et al., 2025).

Yet the human brain goes beyond local, word-by-word prediction. Caucheteux et al. (2023) demonstrated that extending language models with multi-timescale predictions improved their alignment with fMRI data: models trained to anticipate not only the next word but also upcoming words or sentence-level features better matched neural activity, particularly in higher-order frontal and parietal regions that integrate broader context. This suggests that alignment improves when models adopt a predictive coding–like architecture, anticipating information further into the future and at multiple levels of abstraction. dLLMs provide a natural test of this idea: by iteratively refining an entire sequence, they inherently generate predictions with a broader temporal and structural horizon than next-word generators. This iterative, global approach might be a better parallel for how we plan utterances, a process hard to emulate with purely left-to-right generation.

# 3 METHODS

## 3.1 EXTRACTING ECOG DATA DURING SPEECH PRODUCTION

Our ECoG data were drawn from a previously published study (Goldstein et al., 2025) comprising continuous 24/7 recordings from four patients (see Table 2 in Appendix A for a comprehensive description of the patient demographics and clinical characteristics) who engaged in spontaneous conversations with family, friends, doctors, and hospital staff during their week-long stay in the epilepsy monitoring unit. Across patients, neural signals were collected from 675 intracranial electrodes. We selected 466 electrodes located in six left-hemisphere regions of interest (ROIs) defined by the "HCPMMP1_combined" atlas (Glasser et al., 2016): superior temporal gyrus (STG: 100 electrodes), middle and inferior temporal lobe (MTL/ITL: 89 electrodes), inferior frontal gyrus (IFG: 84 electrodes), dorsolateral prefrontal cortex (DLPFC: 55 electrodes), motor cortex (MC: 97 electrodes) and angular gyrus / temporo-parietal-occipital junction (AG/TPOJ: 41 electrodes). These regions have been shown to play critical roles in language use (Malik-Moraleda et al., 2022). All conversations were transcribed, and each word was time-aligned with the concurrent ECoG signals. After preprocessing (see Appendix C for details), we divided the dataset into comprehension (listening) and production (speaking) periods, yielding 50 hours (289,971 words) of comprehension data and 50 hours (230,238 words) of production data in naturalistic settings. For the present study, we focused on production and analyzed 7,229 utterances between 5–25 words in length, excluding short or long productions (see Figure 5 in Appendix B for summary statistics of the sentence length distribution). To accommodate variable utterance lengths, we sampled ECoG activity at five evenly spaced points within each utterance, together with five points from the two seconds preceding and following production, yielding a 15-timepoint ECoG time course for each utterance (see Figure 1).

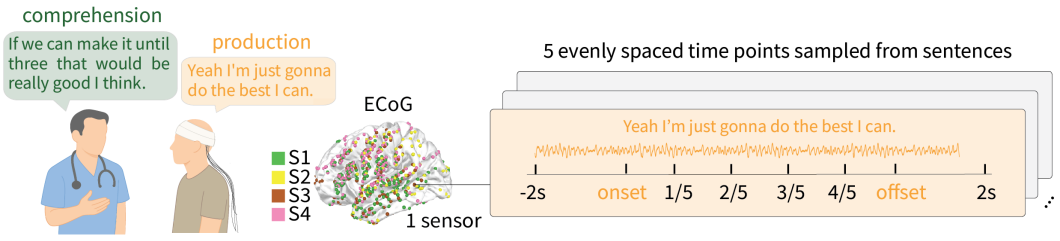


Figure 1: Extracting ECoG data during speech production. ECoG activity at five evenly spaced points within each utterance were sampled, together with five evenly-spaced points from the two seconds preceding and following production, yielding a 15-timepoint ECoG time course for each utterance.

### 3.2 EXTRACTING SENTENCE EMBEDDINGS ACROSS STEPS

We selected LLaDA-8B (Nie et al., 2025) and Dream-7B (Ye et al., 2025) to test their alignment with brain activity during speech production. We also included their autoregressive counterparts LLaMA3-8B (Grattafiori et al., 2024) and Qwen2.5-7B (Yang et al., 2024) for comparison. We extracted sentence embeddings from the 20th layer of the four LLMs across five progressive generation steps. LLaMA3-8B has 32 layers and Qwen2.5-7B has 28 layers in total. The 20th layer was chosen because prior work indicates that representations at approximately two-thirds of a model's depth show the strongest correspondence with neural activity during language processing (Caucheteux & King, 2022). The five-step sampling captured how sentence representations evolve under the distinct generative dynamics of diffusion and autoregressive models. Specifically, for aLLMs (LLaMA, Qwen), we simulated incremental left-to-right generation. Five progressive stages were defined by including the first 20%, 40%, 60%, 80%, and 100% of words in a sentence, providing a linear accumulation of evidence toward the complete sentence (see Figure 2a, upper panel). For the dLLMs (Dream, LLaDA), we implemented an iterative confidence-driven procedure to establish the order in which tokens of a sentence are revealed. This approach differs fundamentally from the left-to-right accumulation of aLLMs: instead of fixing a sequential order a priori, the model dynamically selects the most predictable positions based on context at each iteration (see Figure 2a). We outline the algorithm used to extract sentence embeddings across generative steps from dLLMs:

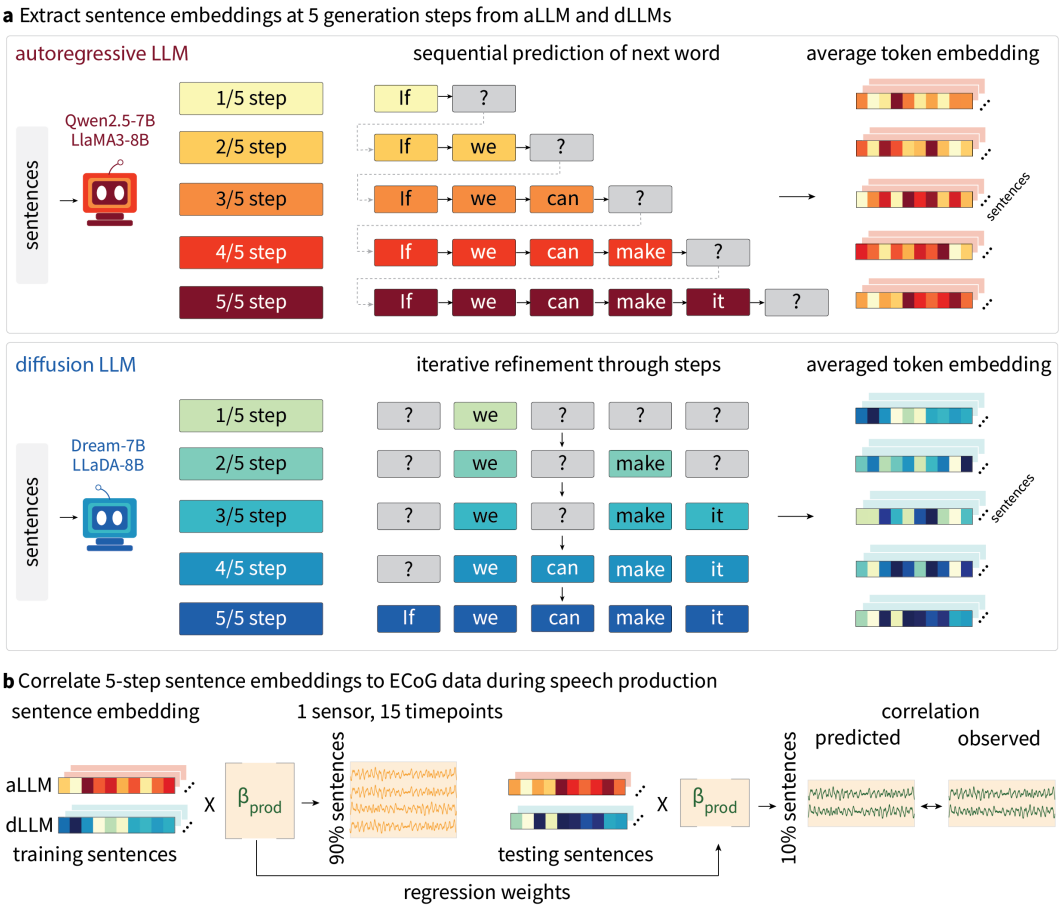


Figure 2: **a**, Extracting sentence embeddings at 5 generation steps in diffusion and autoregressive LLMs. **b**, Aligning aLLM and dLLM sentence embeddings to ECoG activity during speech production using ridge regression.

**Input preparation**. Each target sentence (the production part) was embedded within a conversational prompt (e.g., "In a casual conversation, you heard 'comprehension' and you responded 'production'"). The comprehension field was populated with the preceding utterance from another speaker (see Appendix D for examples). Model forward passes were then executed, and hidden representations from the 20th layer were extracted for the target tokens in the production portion only. To initialize the diffusion process, all target tokens were replaced with a designated mask symbol ([gMASK] for LLaDA and ¡|mask|¿) yielding a fully masked response sequence appended to the prompt.

**Initial position selection.** The fully masked sequence was passed through the model. For every masked position, we computed the softmax probability of the correct target token. The position with the highest confidence score was selected as the first revealed token. This ensured that the diffusion process started from the location where the model was most certain of the ground-truth content given only contextual cues.

**Iterative revelation.** Following this initialization, the model iteratively revealed one additional token per step. At each iteration, previously revealed tokens were fixed in place, while unrevealed positions remained masked. A forward pass generated logits for all masked positions. For each position, the probability assigned to its correct token was extracted. The position with the highest confidence was then revealed and added to the growing set of fixed tokens. Thus, each diffusion model sentence embedding at Step $k$ represents a state where the model has confidently placed $k$ out of $n$ words in their positions, while the remaining words are still masked. This greedy loop continued until all tokens in the target sentence had been revealed.

**Order recording.** The full revelation sequence was stored as an ordered list of positions, where each index indicated the step at which a token was revealed. In practice, this list was aligned to the original tokenization, such that each word could be assigned to the earliest step among its constituent tokens. This produced a word-level revelation trajectory reflecting the model's progressive reconstruction dynamics. Based on the diffusion sequence, each sentence was divided into five steps—20%, 40%, 60%, 80%, and 100% of words (see Figure 2a, lower panel).

**Greedy confidence principle.** The algorithm implements a greedy search strategy: at each step, the most confident masked position is revealed, with no backtracking. Although not globally optimal, this procedure is computationally efficient and reflects a psychologically plausible mechanism of iterative refinement under uncertainty. The algorithm for token revelation in dLLMs is summarized below:

**Algorithm 1** Greedy confidence-based token revelation in dLLMs

**Require:** Model $M$, Tokenizer $T$, Context $c$, Target $s$, Mask token id $m$
**Ensure:** Original tokens, Revelation order, Step indices

Construct prompt
$target\_ids \leftarrow T.encode(s)$
$L \leftarrow$ length($target\_ids$)
$current\_resp \leftarrow [m]^L$
Run $M$ on prompt + $current\_resp$
Compute $conf[p] =$ softmax($logits[p]$)$[target\_ids[p]]$ for all $p$
$best\_pos \leftarrow \arg\max_p conf[p]$
Reveal $best\_pos$; update state and record order
**while** $|revealed| < L$ **do**
    Run $M$ on prompt + $current\_resp$
    **for** each $p$ not in revealed **do**
        $conf[p] \leftarrow$ softmax($logits[p]$)$[target\_ids[p]]$
    **end for**
    $best\_pos \leftarrow \arg\max conf[p]$
    Reveal $best\_pos$; update state and record order
**end while**
**return** original tokens, revelation order, step indices

**Sentence embeddings extraction** After determining the revelation order of words in dLLMs, we computed sentence embeddings for each of the five steps by averaging token-level states at that

step, yielding a single vector per step (dimension 3,584 for Dream and Qwen, 4,096 for LLaDA and LLaMA). The resulting arrays were stored in (n sentences) $\times$ (5 steps) $\times$ (embedding dimension).

### 3.3 Aligning sentence embeddings with ECoG activity

We modeled neural responses from multiple embedding sources using a banded ridge (multiple-kernel) regression (Dupré la Tour et al., 2022), implemented with Himalaya's `MultipleKernelRidgeCV` (see Figure 2b). Each embedding model (Dream, LLaDA, Qwen, LLaMA) was treated as a separate kernel with its own regularization parameter, enabling joint integration of embedding spaces while adaptively weighting their contributions.

Neural responses $\mathbf{y}$ were predicted as $\hat{\mathbf{y}} = \sum_i K_i w_i$, where $K_i = X_i X_i^\top$ and $w_i$ are kernel weights. The regression minimized $\|\mathbf{y} - \sum_i K_i w_i\|^2 + \sum_i \alpha_i w_i^\top K_i w_i$, with independent ridge penalties $\alpha_i$ per kernel. We used the `precomputed` kernel option with random search over $\alpha_i \in [10^0, 10^{20}]$, optimizing log-weights $\delta_i = -\log \alpha_i$ via cross-validation. Data were split 90%/10% into training and testing sets in temporal order. Per-kernel predictions $\hat{\mathbf{y}}_i$ were obtained using `predict(split=True)`, and Pearson correlations with observed responses were computed as model-specific scores. This was repeated for each embedding step (1–5) and timepoint, producing a tensor of size ($5 \times 15 \times 4$ models).

Statistical significance was assessed with non-parametric permutation tests. Correlation scores were aggregated across all electrodes, with LLaDA+Dream averaged as diffusion and LLaMA+Qwen as autoregressive. Null distributions were generated from 1000 random permutations across 50,625 comparisons (5 steps $\times$ n electrodes $\times$ 15 timepoints), and $p$-values were computed as the proportion of permuted values exceeding the observed score.

### 3.4 Analyzing word-level features across different generation steps

**Visualizing embeddings with PCA.** We applied principal component analysis (PCA) to layer-20 embeddings from dLLMs and aLLMs to visualize how sentence representations evolve across five steps. For each subject, embeddings of shape (n sentences $\times$ 5 steps $\times$ dimension) were concatenated and reshaped so each sentence–step pair was treated as a data point. PCA reduced dimensionality to three components, performed separately for diffusion and autoregressive LLMs, with explained variance averaged across models.

**Measuring diffusion–autoregressive distances.** Representational differences were quantified using Jensen–Shannon (JS) divergence between 5-step embeddings from LLaMA vs. LLaDA and Qwen vs. Dream. At each step, sentence-level JS distances were computed for paired autoregressive-diffusion LLMs and averaged across the two pairs.

**Word frequency across steps.** Log word frequencies for words at each step were retrieved from the Google Books N-gram corpus (Google, 2010) for both model families and compared across steps using paired $t$-tests with FDR correction for multiple comparisons.

**POS distributions across steps.** Part-of-speech (POS) tags obtained with spaCy were grouped into four categories: NOUN (including PROPN, PRON), VERB (including AUX), ADJ/ADV, and FUNC (all remaining tags). For each generation step, tag proportions were computed relative to the sentence total, averaged across Dream and LLaDA, and compared against autoregressive distributions using paired $t$-tests with FDR correction for multiple comparisons.

## 4 Results

### 4.1 Word-level features across diffusion and autoregressive steps

Table 1 presents five illustrative examples of words generated across the five diffusion steps. At Step 1 the most predictable token (often a content word like "dinner" or "her husband") is placed, by Step 5 the full sentence is formed. Figure 3a shows the first three PCs of the 5-step embeddings for aLLMs and dLLMs. dLLMs exhibited a clearer temporal trajectory than aLLMs: embeddings from successive steps were more distinctly separated in principal component space. Moreover, the first PC of diffusion embeddings explained substantially more variance (25.2%) than that of autoregressive

embeddings (9.7%), highlighting stronger step-wise differentiation in dLLMs. We further quantified these differences by computing Jensen–Shannon divergence between model families (Figure 3b), which revealed consistently greater separation at earlier steps.

Significant differences also emerged in the distribution of log word frequency across generative steps (Figure 3c). At the onset and offset of sentences, dLLMs produced words of significantly higher frequency than aLLMs (Step 1: $t = 16.21$, $p < 10^6$; Step 5: $t = 17.16$, $p < 10^6$). In contrast, in the middle portions of sentences, aLLMs favored higher frequency words (Step 3: $t = -10.74$, $p < 10^6$; Step 4: $t = -13.66$, $p < 10^6$). No significant difference was observed at Step 2 ($t = 1.86$, $p = 0.063$). These results reveal a U-shaped frequency pattern for dLLMs—relying on common, high-frequency words at sentence onset and offset while using relatively lower frequency words mid-sentence—whereas aLLMs show the opposite tendency in the middle steps.

Finally, POS distributions diverged systematically between model types (Figure 3d): Paired $t$-tests reveal consistent boundary effects: dLLMs favored more content words (NOUN, VERB) at Step 1 ($t = 6.08$, $p < 10^6$) and Step 4 ($t = 2.40$, $p = 0.017$), whereas aLLMs used more NOUNs at Step 5 ($t = -18.92$, $p < 10^6$). VERB and FUNC categories likewise exhibited mirrored trends: dLLMs used more verbs at sentence start (Step 1: $t = 13.18$, $p < 10^6$), while aLLMs dominated the later steps. ADJ/ADV showed smaller but significant differences in a few steps. The shifting POS proportions support the hypothesis that dLLMs reorganize lexical categories across the course of generation differently than aLLMs.

Table 1: Examples of original sentences and reordered outputs from LLaDA and Dream across five generation steps.

| Source | Step 1 | Step 2 | Step 3 | Step 4 | Step 5 |
|---|---|---|---|---|---|
| Original | Uh where | are they | going | for | dinner |
| LLaDA | dinner Uh | where they | are | going | for |
| Dream | dinner Uh | where they | are | going | for |
| Original | Mhm I | see my | sister and | her | husband |
| LLaDA | her husband | Mhm see | I my | sister | and |
| Dream | her husband | Mhm see | sister I | my | and |
| Original | Yeah Eh | it'll give | her | some | cushion |
| LLaDA | Eh cushion | Yeah it'll | give | her | some |
| Dream | her cushion | Yeah it'll | Eh | give | some |
| Original | We'll | try | our | best | together |
| LLaDA | try | We'll | our | best | together |
| Dream | try | We'll | best | our | together |
| Original | Not a | big soda | I don't | drink | soda |
| LLaDA | soda I | drink soda | Not big | a | don't |
| Dream | big soda | drink soda | Not I | a | don't |

### 4.2 Brain encoding performance

Figure 4 illustrates the encoding performance of aLLMs and dLLMs during naturalistic speech production. Encoding performance was quantified as the correlation coefficient between predicted and observed ECoG activity, evaluated across five generative steps and six cortical ROIs: STG, MTL+ITL, IFG, DLPFC, MC, AG/TPOJ.

**Autoregressive LLMs.** Time-resolved encoding (panel a) reveals modest correlations across ROIs, with encoding performance gradually increasing after sentence onset and peaking around sentence offset. Although correlations are generally weak, later generative steps (Steps 4–5, shown in darker red) consistently outperform earlier steps (Step 1, light yellow), suggesting that autoregressive embeddings become more predictive of neural activity as sentence production unfolds. The corresponding heatmaps (panel b) confirm this trend: across all ROIs, Step 5 embeddings yield the strongest correlations, particularly in IFG and MC, while Step 1 embeddings show weaker encoding. We

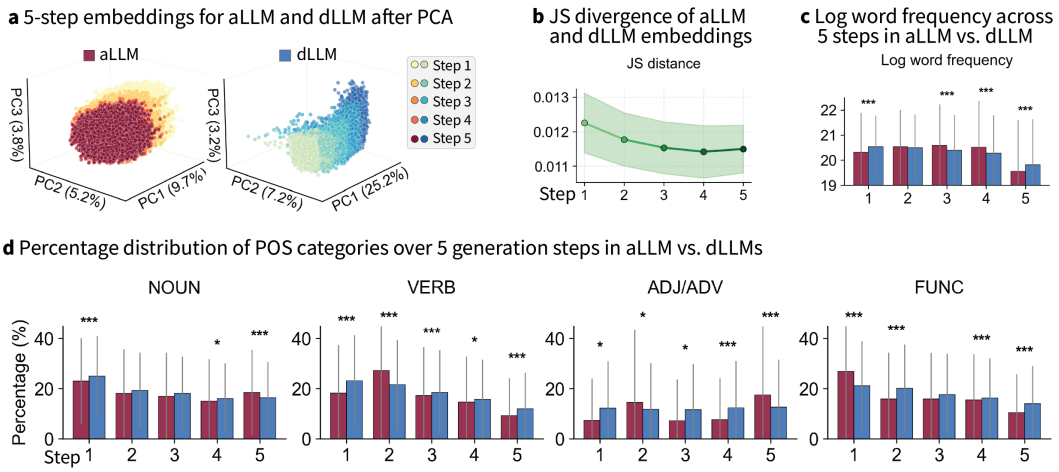


Figure 3: Structural and lexical differences across generative steps in aLLMs and dLLMs. **a** PCA shows greater step-wise separation in diffusion embeddings than in autoregressive ones. **b** Jensen–Shannon divergence decreases across steps. **c** dLLMs favor higher-frequency words at sentence boundaries. **d** dLLMs produce more content words early, whereas aLLMs favor function words.

also evaluated LLaMA and Qwen separately and found no meaningful difference in their encoding performance (see Figure 6 in Appendix E).

**Diffusion LLMs.** dLLM embeddings exhibit a systematic temporal alignment with neural dynamics during speech production. Earlier diffusion steps (Steps 1–2) correlated more strongly with neural activity in temporal regions such as STG and MTL+ITL and DLPFC prior to articulation, suggesting that these representations capture pre-articulatory planning processes. By contrast, later diffusion steps (Steps 4–5) achieved higher correlations in STG, IFG, DLPFC, MC, and AG/TPOJ around and after sentence offset, consistent with mid- and post-articulatory stages of production (panel c, d). This step-specific mapping was less pronounced in aLLMs, where encoding performance increased more gradually across steps. Together, these findings indicate that dLLMs not only capture overall neural dynamics but also differentiate between pre-articulatory planning and later motor-related processes across cortical regions. We also analyzed LLaDA and Dream separately, and although some differences were observed in the motor cortex, our main conclusions remain unchanged (see Figure 7 in Appendix E).

## 5 Discussion and Conclusion

This work provides the first direct evidence that dLLMs capture neural dynamics of human speech production in ways that qualitatively differ from aLLMs. Whereas aLLMs gradually increase neural predictivity as tokens accumulate, dLLMs show sharper step-wise differentiation: earlier diffusion steps align with pre-articulatory activity while later steps align with activity during and after articulation within middle/inferior temporal and motor cortices. These results suggest that the brain's production system may function more like an iterative refinement process than a strictly left-to-right generator. From a cognitive neuroscience perspective, this refinement offers a mechanism for balancing early commitments (e.g., reliance on high-frequency words at utterance boundaries) with later flexibility in lexical and syntactic choice, thereby reconciling evidence for both sequential planning and global message-level representations in speech production.

Nevertheless, several caveats remain: We only examined a single layer (layer 20) from each model; it remains possible that other layers, or combinations of layers, exhibit stronger or qualitatively different alignment. Moreover, our token revelation procedure reflects one particular instantiation of diffusion dynamics; alternative denoising strategies may produce different mappings. Future work

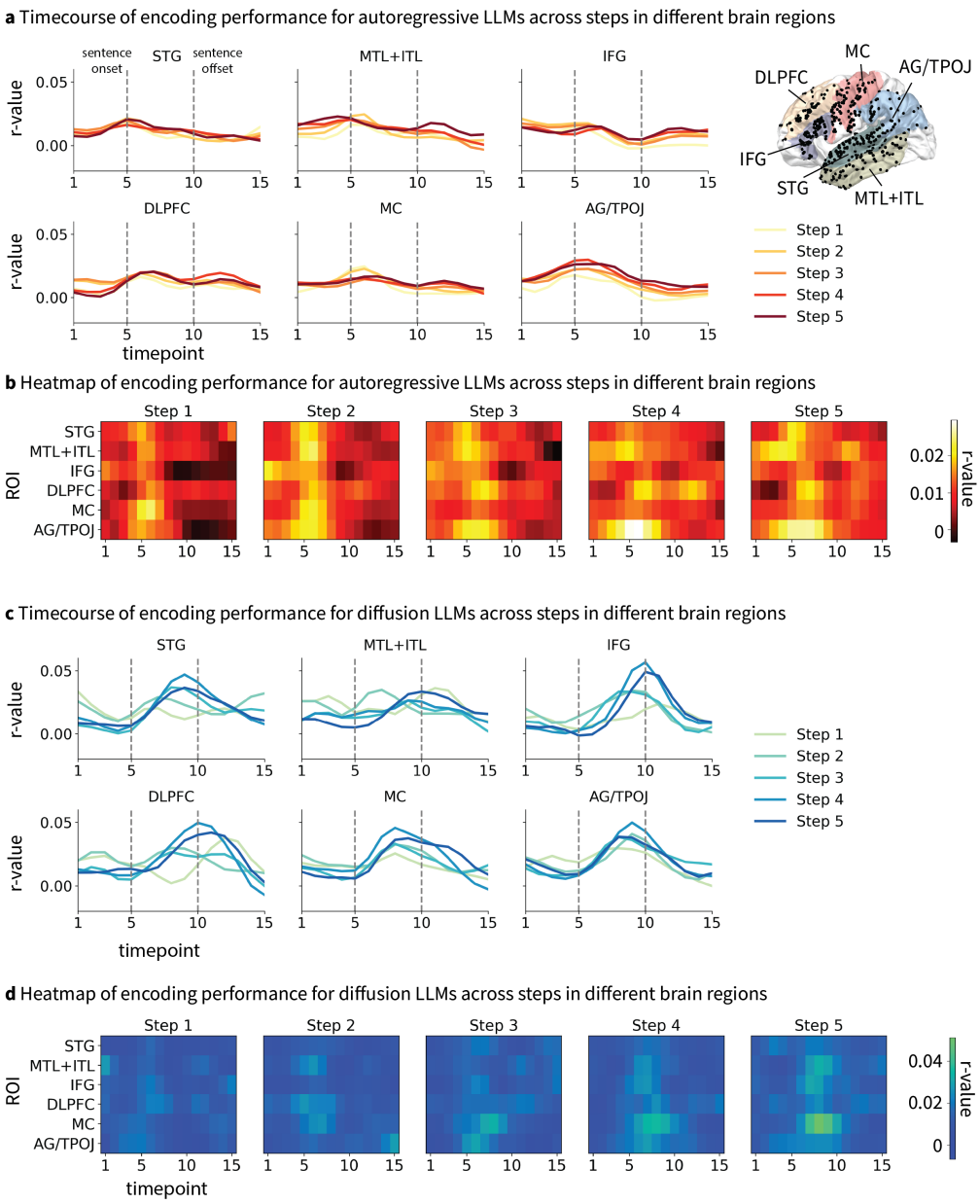


Figure 4: Encoding performance of aLLMs and dLLMs during speech production. **a,b** aLLMs show gradual increases in encoding performance across steps, with later steps aligning more strongly around articulation. **c,d** dLLMs exhibit sharper step-wise differences: early steps correlate with pre-articulatory activity in temporal regions, while later steps align with mid- and post-articulatory activity in frontal and motor regions.

should therefore expand the range of models, datasets, and tasks considered, and test whether similar dynamics hold across modalities such as fMRI, MEG, or in cross-linguistic production.

In conclusion, our study shows that dLLMs do more than merely rival aLLMs on linguistic benchmarks — they offer a cognitively plausible model of human speech production. This work opens several avenues: testing dLLM–brain alignment with other modalities (fMRI, MEG), exploring finer-grained layer-wise dynamics, and developing hybrid models that integrate sequential and diffusion principles.

## ETHICS STATEMENT

Participants gave informed consent in accordance with protocols approved by the XXX University Institutional Review Board. They were explicitly informed that participation was independent of their clinical care and that withdrawal would not affect their medical treatment. We do not anticipate any additional ethical concerns.

## REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the procedures for extracting sentence embeddings and implementing banded ridge regression encoding models. All computations were performed using standard Python packages, including Hugging Face for language model handling and MNE-Python for neural data processing. The code and associated data for replicating the encoding analyses will be released publicly upon acceptance of the paper.

## REFERENCES

Richard Antonello, Aditya Vaidya, and Alexander G. Huth. Scaling laws for language encoding models in fmri. In *Advances in Neural Information Processing Systems*, 2024.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems*, 2021.

Kristofer E. Bouchard, Nima Mesgarani, Keith Johnson, and Edward F. Chang. Functional organization of human sensorimotor cortex for speech articulation. *Nature*, 495(7441):327–332, 2013.

Paul Broca. Remarques Sur le Siége de la Faculté Du Langage Articulé, Suivies D'une Observation D'aphémie (Perte de la Parole). *Bulletin de la Société Sciences Nat*, 6:330–357, 1861.

Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1), 2022.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature Human Behaviour*, 7(3):430–441, 2023. doi: 10.1038/s41562-022-01516-2.

Deepmind. Gemini diffusion, 2024. URL `https://deepmind.google/models/gemini-diffusion`.

Gary Dell. A spreading-activation theory of retrieval in sentence production. *Psychological Review*, 93(3):283–321, 1986. doi: 10.1037/0033-295X.93.3.283.

Gary Dell. The retrieval of phonological forms in production: Tests of predictions from a connectionist model. *Journal of Memory and Language*, 27(2):124–142, 1988. doi: 10.1016/0749-596X(88)90070-8.

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

Changjiang Gao, Zhengwu Ma, Jiajun Chen, Ping Li, Shujian Huang, and Jixing Li. Increasing alignment of large language models with language processing in the human brain. *Nature Computational Science*, 2025. doi: 10.1038/s43588-025-00863-0.

Merrill F. Garrett. Syntactic process in sentence production. In *Psychology of learning and motivation: Advances in research and theory*, pp. 133–177. 1975.

Merrill F. Garrett. The limits of accommodation. In *Errors in linguistic performance*, pp. 263–271. New York: Academic, 1980.

Matthew F. Glasser, Timothy S. Coalson, Emma C. Robinson, Carl D. Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F. Beckmann, Mark Jenkinson, Stephen M. Smith, and David C. Van Essen. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel 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.

Ariel Goldstein, Haocheng Wang, Leonard Niekerken, Mariano Schain, Zaid Zada, Bobbi Aubrey, Tom Sheffer, Samuel Nastase, Harshvardhan Gazula, Aditi Singh, Aditi Rao, Gina Choe, Catherine Kim, Werner Doyle, Daniel Friedman, Sasha Devore, Patricia Dugan, Avinatan Hassidim, Michael Brenner, Yossi Matias, Orrin Devinsky, Adeen Flinker, and Uri Hasson. A unified acoustic-to-speech-to-language embedding space captures the neural basis of natural language processing in everyday conversations. *Nature Human Behavior*, 9(5):1041–1055, 2025.

Google. Google books ngram viewer. `https://books.google.com/ngrams/`, 2010. Accessed: 2025-09-25.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh,

Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Zhuoqiao Hong, Haocheng Wang, Zaid Zada, Harshvardhan Gazula, David Turner, Bobbi Aubrey, Leonard Niekerken, Werner Doyle, Sasha Devore, Patricia Dugan, Daniel Friedman, Orrin Devinsky, Adeen Flinker, Uri Hasson, Samuel A Nastase, and Ariel Goldstein. Scale matters: Large language models with billions (rather than millions) of parameters better match neural represen-

tations of natural language. *eLife*, 2024. doi: 10.7554/elife.101204.1.

Shailee Jain and Alexander Huth. Incorporating context into language encoding models for fmri. In *Advances in Neural Information Processing Systems*, 2018.

Inception Labs, Samar Khanna, Siddhant Kharbanda, Shufan Li, Harshit Varma, Eric Wang, Sawyer Birnbaum, Ziyang Luo, Yanis Miraoui, Akash Palrecha, Stefano Ermon, Aditya Grover, and Volodymyr Kuleshov. Mercury: Ultra-fast language models based on diffusion, 2025.

Willem Levelt. *Speaking: From intention to articulation*. MIT Press, Cambridge, MA, 1989. ISBN 9780262121274.

Willem Levelt. Producing spoken language: A blueprint of the speaker. In *The neurocognition of language*, pp. 83–122. Oxford University Press, 1999.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B. Hashimoto. Diffusion-lm improves controllable text generation. In *Advances in Neural Information Processing Systems*, 2022.

Saima Malik-Moraleda, Dima Ayyash, Jeanne Gallée, Josef Affourtit, Malte Hoffmann, Zachary Mineroff, Olessia Jouravlev, and Evelina Fedorenko. An investigation across 45 languages and 12 language families reveals a universal language network. *Nature Neuroscience*, 25(8):1014–1019, 2022.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models, 2025. URL https://arxiv.org/abs/2502.09992.

Ned T. Sahin, Steven Pinker, Sydney S. Cash, Donald Schomer, and Eric Halgren. Sequential processing of lexical, grammatical, and phonological information within broca's area. *Science*, 326(5951):445–449, October 2009.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021.

Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Advances in Neural Information Processing Systems*, 2019.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b, 2025. URL https://hkunlp.github.io/blog/2025/dream/.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022.

## A PARTICIPANTS

Table 2: Patient demographics and clinical characteristics.

| | **P1** | **P2** | **P3** | **P4** |
|---|---|---|---|---|
| Age (years) | 53 | 26 | 48 | 24 |
| Sex | F | M | F | M |
| Electrodes implanted | 104 | 125 | 255 | 192 |
| Hours of speech recorded | 17 | 37 | 17 | 29 |
| Words recorded | 79,654 | 213,473 | 117,800 | 109,282 |
| Comprehension words | 47,642 | 109,967 | 71,754 | 60,608 |
| Production words | 32,012 | 103,506 | 46,046 | 48,674 |
| Pathology / Seizure focus | Posterior temporal lobe (neocortical) epilepsy; seizure focus adjacent to posterior temporal lesion | Left anteromedial temporal lobe epilepsy | Right anteromedial temporal lobe epilepsy; ictal onsets localized to temporal pole and hippocampus | Focal epilepsy in left hemisphere with broad focus including temporal neocortex, frontal operculum, postcentral gyrus, insula |
| Implant | Left grid, strips, depth | Left grid, strips, depth | Bilateral strips/depths, left grid | Left grid, strips, depth |

## B SENTENCE DATA

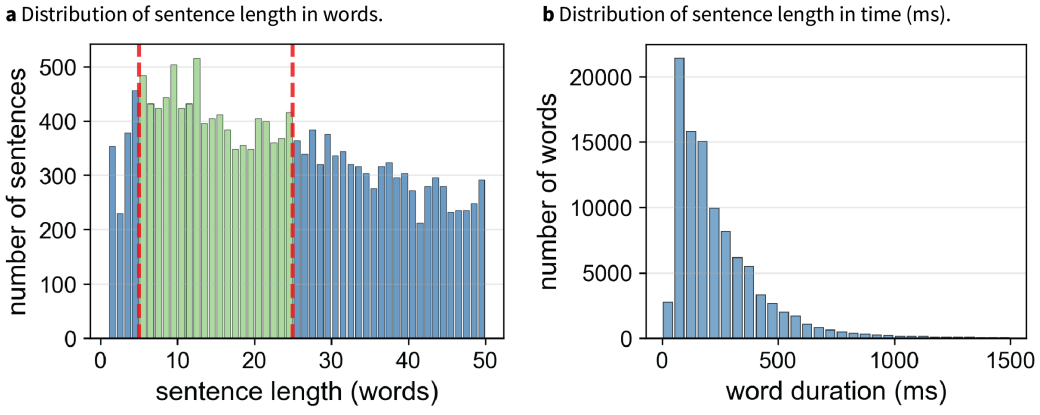


Figure 5: Summary statistics of production sentences. **a** Distribution of sentence length in words. Dashed lines indicate the 5-25 words threshold. **b** Distribution of sentence length in time (ms).

## C ECOG PREPROCESSING

The ECoG preprocessing pipeline mitigated artifacts arising from movement, faulty electrodes, line noise, abnormal physiological signals (e.g., epileptic discharges), eye blinks, and cardiac activity. A semi-automated procedure was used to identify and remove corrupted data segments (e.g., seizures, loose wires), while additional noise was attenuated using fast Fourier transform (FFT), independent

component analysis (ICA), and de-spiking methods. Neural signals were then band-pass filtered in the broadband range (75–200 Hz), and the power envelope was computed as a proxy for each electrode's average local firing rate. The resulting signals were z-scored, smoothed with a 50-ms Hamming kernel, and trimmed by 3,000 samples at each end to minimize edge effects. All preprocessing was conducted using custom MATLAB 2019a (MathWorks) scripts.

## D PROMPT EXAMPLES

**Conversation:**
Nurse (comprehension): `"Oh my hands are cold."`
Patient (production): `"Oh they do actually."`

**Prompt to LLMs:**
`In a casual conversation, you heard "Oh my hands are cold." and you responded "Oh they do actually."`

**Conversation:**
Nurse (comprehension): `"You're not gonna need it in two days. It's temporary."`
Patient (production): `"I know, but I feel like in two days they're probably gonna go take this thing out of my head."`

**Prompt to LLMs:**
`In a casual conversation, you heard "You're not gonna need it in two days. It's temporary." and you responded "I know, but I feel like in two days they're probably gonna go take this thing out of my head."`

## E BRIAN ENCODING PERFORMANCE OF EACH LLM.

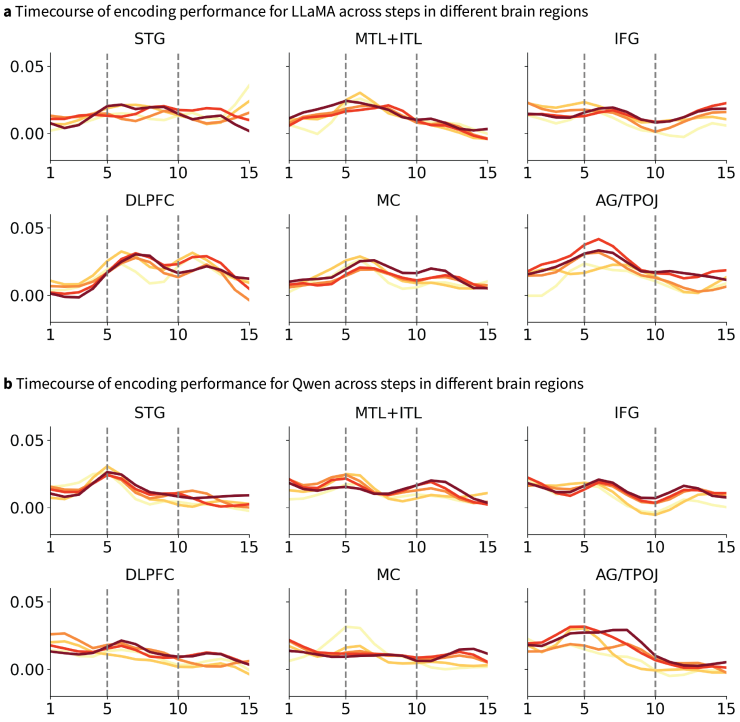


Figure 6: Timecourse of encoding performance for LLaMA (**a**) and Qwen (**b**) across steps in different brain regions.

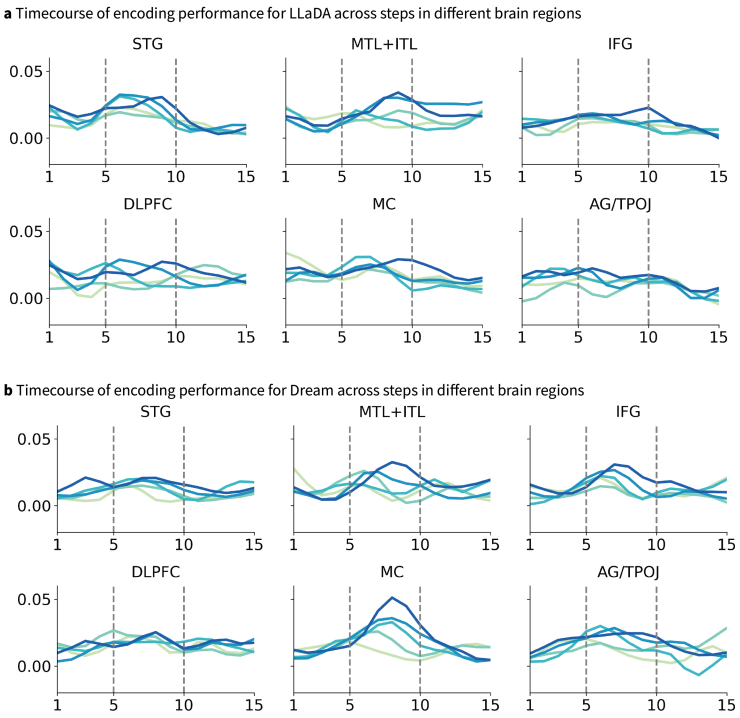


Figure 7: Timecourse of encoding performance for LLaDA (**a**) and Dream (**b**) across steps in different brain regions.

## F Implementation details

All computations for extracting dLLM and aLLM sentence embeddings and for running ridge regression brain-encoding analyses were performed on a high-performance computing (HPC) cluster equipped with 128 CPU cores and two A100 GPUs per node.