BINDING PHONEME IDENTITY TO POSITION IN NEURAL MODELS AND THE HUMAN BRAIN

Anonymous authors
Paper under double-blind review

ABSTRACT

A fundamental challenge in cognitive science and neural computation is the binding problem: how neural systems associate separately represented contents with their structural roles. Compositional generalization can be achieved by factorizing representations into *roles*, the structural slots of a representation (e.g., serial position or syntactic function), and *fillers*, the contents that occupy them (e.g., a phoneme or a word). Because roles and fillers can be recombined freely, a system that represents them separately can represent novel combinations. How neural systems implement role–filler binding, however, remains largely unknown. One of its simplest linguistic cases is phoneme sequencing, where phoneme identity (the filler) must be bound to position (the role) so that /gum/ and /mug/ are represented differently. While previous work identified the neural encoding of single phonemes in human auditory cortex, it remains unclear how these encodings are combined into sequences. A recent proposal from natural language processing, the 'onion hypothesis' (Csordás et al., 2024), offers a concrete account: the direction of the neural population state encodes token identity, while its magnitude encodes token position. However, the onion hypothesis was formulated for random token sequences; here we extend it to natural language, where successive tokens are strongly constrained, and test it in both neural models and the human brain. We first evaluated the onion hypothesis in an encoder-decoder trained to repeat English words, combining representational geometry analyses with causal interventions. Magnitude alone turns out to be incomplete: position is encoded by magnitude together with a systematic rotation of the state direction, a scheme we formalize as the 'spiral hypothesis'. We then analyzed intracranial recordings from the auditory cortex of 20 participants listening to short spoken sentences. Across superior temporal gyrus electrodes, the change in population response between consecutive phonemes of a word decreases in magnitude as position advances, while the direction of that change is positively aligned within consonants and within vowels and close to orthogonal across the two classes. This supports the magnitude coding shared by both the onion and spiral accounts. Together, these results provide a geometric account of how neural systems bind phoneme identity to position, and show how hypotheses derived from neural models can yield testable predictions about human neural computation.

1 INTRODUCTION

Generalizing beyond the training distribution remains a central difficulty for AI models. When carefully tested, AI models still do so less systematically than humans, in, e.g., reasoning (Wu et al., 2024; Chollet et al., 2026), vision (Campbell et al., 2024), and language (Lake & Baroni, 2018; Ismayilzada et al., 2025). One explanation for this gap is that humans better extract and disentangle stable structures in the data from their contents. In cognitive science, this distinction is formalized as one between *roles*, the structural slots of a representation (e.g., a location in a spatial layout, or a serial position in a sequence), and *fillers*, the contents that occupy them (e.g., the landmark at that location, or the phoneme at that position) (Smolensky, 1990). A system that represents roles and fillers separately can reuse a familiar role with new fillers, rather than memorizing each combination on its own (Roskies, 1999; Greff et al., 2020). Such role–filler factorization is useful across many domains, such as spatial navigation (Behrens et al., 2018) and language (Rambaud et al., 2026; Feng

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

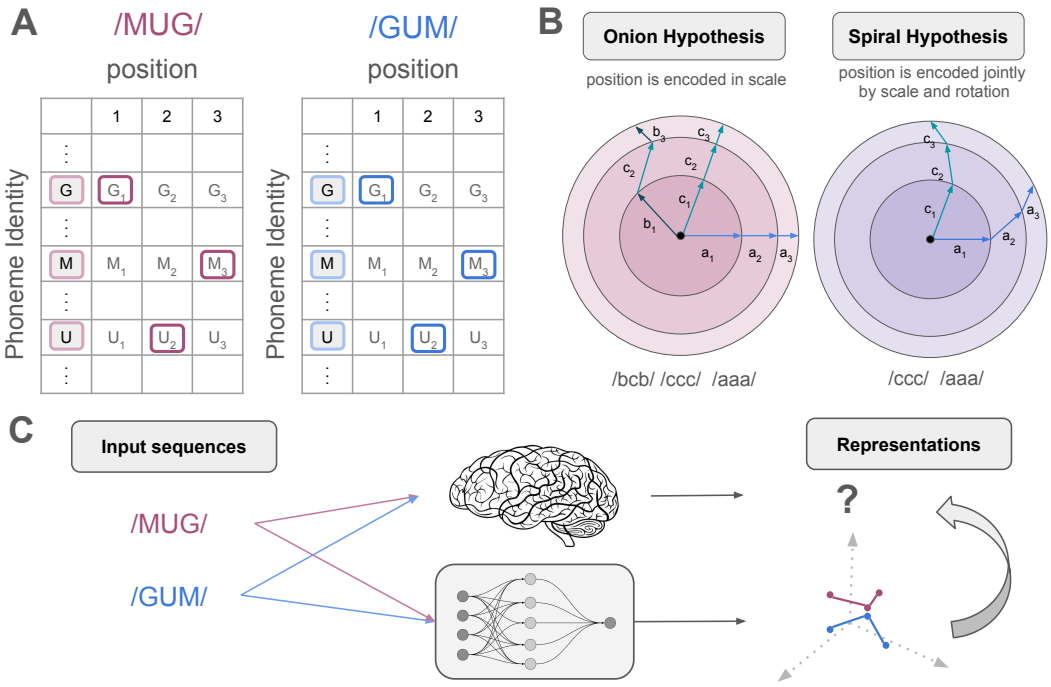


Figure 1: **Hypotheses for How a Recurrent State Binds Phoneme Identity to Position.** **(A)** The binding problem for sequences. Two words made of the same phonemes in a different order, /gum/ and /mug/, must map to different states, so the representation cannot depend on phoneme identity alone (e.g., a bag-of-phoneme model). **(B)** Two hypotheses for how position is encoded. Under the onion hypothesis (left) identity is carried by the direction of the representation and position by its magnitude, so a repeated phoneme keeps its direction across positions while its magnitude shrinks. Under the spiral hypothesis (right) identity is again carried by direction, but position is carried jointly by magnitude and rotation. **(C)** A recurrent network trained to repeat words serves as a hypothesis generator. Its positional representation can be both measured and causally manipulated, and the predictions it yields are then tested in human auditory-cortex recordings.

& Steinhardt, 2024). How neural systems implement this factorization, and the binding it requires, remains an open question.

Here we take the neural encoding of phoneme sequences as a case study. It is simple enough to make the geometry measurable: the phoneme inventory is small and discrete, and words are only a few phonemes long. Furthermore, the encoding of individual phonemes is already well characterized, in human superior temporal gyrus (STG) as combinations of phonetic features (Mesgarani et al., 2014a; Yi et al., 2019). This prior work characterizes the encoding of identity alone, however, leaving open how it is bound to a position.

To address this question, we build on a hypothesis recently proposed in natural language processing for recurrent neural networks (RNNs). RNNs were inspired by cortical recurrence (Elman, 1990; Hopfield, 1982). Unlike transformers, which attend over a stored history of past inputs, RNNs update a fixed-size state as each input arrives, a mechanism closer to how the brain processes sequences. This also gives them a computational advantage: they process a sequence in linear time and at constant memory per step, avoiding the growing cache that attention requires. For this reason recurrent architectures have come back into wide use in the form of state-space models (Gu et al., 2022; Gu & Dao, 2024). In such recursive architecture, identity and position must be jointly encoded in a single evolving state, unlike in transformers, where position is supplied by a separate positional embedding. The *Onion Hypothesis* gives a concrete account of how this is done: the direction of the state encodes token identity and its magnitude encodes token position, with earlier tokens held at larger magnitude (Csordás et al., 2024). However, the onion hypothesis was formulated for random token sequences, whereas natural language is strongly constrained by phonotactics (Vitevitch & Luce, 1999; Leonard

et al., 2015), consonant–vowel alternation (Clements, 1990), and frequency (Piantadosi, 2014). This leaves two open questions: does the onion hypothesis hold for natural language, and does it capture how the human brain binds phoneme identity to position?

We first evaluated the Onion Hypothesis in a pre-trained encoder-decoder RNN trained to repeat English words at natural frequencies (Dager et al., 2025). A hypothesis-free analysis of the encoder geometry reproduced the predicted magnitude decay, but showed that the direction of a phoneme's contribution is not accurately preserved across positions, which the hypothesis requires. We therefore propose a refinement, the *Spiral Hypothesis*, in which position is carried jointly by a decaying magnitude and a systematic rotation of the state direction. To compare the two accounts we turned to causal interventions on the encoder state, editing it as each hypothesis prescribes and asking whether the decoder then produces the word that hypothesis predicts. The spiral model outperformed the onion model and also a subspace baseline. Furthermore, a control experiment confirmed that its advantage comes from the rotation itself rather than from its additional parameters. Interventions also failed more often when the edit would produce a phonotactically illegal word, showing that the positional code is shaped by the statistics of the language – an effect that random sequences cannot reveal.

We then asked whether this geometry is present in the human brain, analyzing intracranial recordings from 20 participants listening to short spoken sentences. Across superior temporal gyrus electrodes, the change in population response between consecutive phonemes decreases in magnitude as position advances, as both hypotheses predict. This decline persists when phoneme predictability is controlled, indicating that magnitude encodes position rather than surprisal. The direction of the change is aligned within consonants and within vowels and close to orthogonal across the two classes, consistent with direction encoding phoneme identity. Together, these results give a geometric account of how phoneme identity is bound to position, and show how a hypothesis developed for neural language models can yield testable predictions about the human brain.

2 Experimental Setup

2.1 Recurrent Model and Contextualized Phoneme Embeddings

Pretrained Model. We study the pretrained Neural Word Repetition (NWR) model of Dager et al. (2025). It is an encoder–decoder network with LSTM units (Hochreiter & Schmidhuber, 1997), trained on a word repetition task. The encoder and decoder each have a single hidden layer of size $d = 128$, and the input and output vocabulary consists of 42 phonemes. The model was trained on the 30,000 most frequent English words, sampled according to their frequency. Its hyperparameters (d, number of layers, etc.) were chosen as the smallest model that achieved perfect repetition on the training words. After training, the NWR generalizes to unseen pseudowords (phoneme sequences outside the training vocabulary that follow English phonotactics), repeating them with 97.25% accuracy (Dager et al., 2025).

Contextualized Phoneme Embeddings. We take the encoder state to be the concatenation $z = [h; c] \in \mathbb{R}^{2d}$ of the hidden and the cell state, since the information the decoder reads may be distributed across both (Appendix F reports the analyses on each component separately). To achieve strong repetition performance, z has to encode every preceding phoneme identity together with its position. To isolate the representation of a single phoneme in the sequence, we compute

$$\Delta z_i \;=\; z_i - z_{i-1}, \qquad z_i = \mathrm{Encoder}(w_{1:i}), \tag{1}$$

for every word w. Since z_i accumulates the encoding of all phonemes up to position i, Δz_i isolates the update induced by the phoneme at position i, given the preceding context. We therefore refer to Δz_i as the *contextualized embedding* of that phoneme, in contrast to the non-contextualized input embedding, which depends on phoneme identity alone. The onion hypothesis makes two predictions about these embeddings: their norm should decrease with position, and their direction should carry phoneme identity and therefore be preserved when the same phoneme occurs at a different position.

Representational Analyses. Two descriptive measures test these predictions. For the positional prediction, we take the ℓ_2 norm of Δz_i as a function of position. For the identity prediction, we compute the cosine similarity between pairs of contextualized embeddings at two levels of granularity.

At the fine level, pairs are grouped by whether they share the same phoneme identity; if direction encodes identity, embeddings of the same phoneme should be more aligned than those of different phonemes. At the coarse level, pairs are grouped by broad phonetic class (consonant or vowel); if direction encodes identity, this structure should carry over to classes, with higher alignment within a class than across classes. The same measures are applied to the human recordings in Section 2.3, where the identity prediction is tested at the coarse level only.

2.2 Causal Interventions

Representational analyses and descriptive geometry show how representations are organized but cannot establish whether a geometric property is used when the model produces an output. For causal claims, we adopt the causal abstraction framework (Geiger et al., 2021): we edit the encoder state so that it should correspond to a sequence in which one phoneme has been replaced, and check whether the frozen decoder generates that counterfactual sequence.

Counterfactual Data. Counterfactual pairs are built from the 30,000 training words by choosing a position at random and replacing its phoneme with one sampled from the vocabulary. Each example thus pairs a real word with a counterfactual that differs from it at a single position p (Appendix Table 5). The encoder reads one member of the pair, the *base*, and the decoder must produce the other, the *target*. We run both directions: taking the counterfactual as target requires the intervention to write an unattested phoneme sequence into the state, and taking the real word as target requires it to undo that edit. We verified that all base and targets are repeated correctly by the frozen model before any intervention, so a failure reflects the edit rather than the model. Because random substitution often yields sequences that no English word could potentially contain, we also define two phonotactic filters over the counterfactuals: one enforces the consonant–vowel patterns seen in training, and the other rejects phoneme bigrams or trigrams absent from the training words. Appendix C describes the filters and varies them as an experimental factor.

Additive Intervention. Following Csordás et al. (2024), we write the change of identity into the state by adding a vector to it. Let x_{old} and x_{new} be the embeddings of the base and target phonemes at position p. A learned *positional scaling* vector $\phi(p) \in \mathbb{R}^{2d}$ controls how the change is written into the state,

$$z' = z + \phi(p) \odot (x_{\text{new}} - x_{\text{old}}), \tag{2}$$

where $\odot$ is elementwise multiplication. The embedding difference sets the direction of the change, and $\phi(p)$ sets how strongly and along which axes that change enters the state at position p. The edited state z' is passed to the frozen decoder, and ϕ is trained by minimizing the cross-entropy between the target sequence and the decoder output, generated without teacher forcing. We report exact-sequence match throughout.

Identity Embeddings. The embeddings x_{old} and x_{new} are initialized in one of three ways: (i) from the input non-contextualized embedding layer of the pretrained model, (ii) from the mean contextualized embedding Δz of each phoneme over the training set (Section 2.1), or (iii) ranodmlly (i.e., learned from scratch). For (i) and (ii), we evaluate fixed embeddings (i.e., no learning) and fine-tuned embeddings, jointly with ϕ.

Positional scaling functions. We parameterize the positional scaling $\phi(p)$ as a function of position, $\phi(p) = f_\theta(p)$, with parameters θ shared across positions. Each choice of f_θ is a structural hypothesis about how position enters the state; because f_θ is defined for every p, it also predicts the scaling at positions on which it was not fitted. We compare three scaling functions, with $8 \times d$, $6 \times d$, and $7 \times d$ learnable parameters, respectively. The first reproduces the original onion account, the second isolates its core claim of magnitude decay, and the third adds a rotation of the identity direction with position. Powers of vectors are taken elementwise.

(i) **Onion.** As a reference, we use the scaling function of Csordás et al. (2024), an exponential decay with a linear drift:

$$\phi(p) = g \odot \gamma^p + \beta\, p + b, \qquad g, \gamma, \beta, b \in \mathbb{R}^{2d}. \tag{3}$$

(ii) **DECAY.** The linear term in ONION is unbounded and can grow with position, which departs from the hypothesis that magnitude decreases with position. DECAY removes this term and constrains the scaling to decay monotonically:

$$\phi(p) = s(p) = \mathrm{softplus}(g) \odot \sigma(\gamma)^p + b, \qquad g, \gamma, b \in \mathbb{R}^{2d}, \tag{4}$$

where σ is the logistic sigmoid. Since $\sigma(\gamma) \in (0,1)^{2d}$ and $\mathrm{softplus}(g) > 0$, each dimension decays monotonically to its baseline b. DECAY thus tests whether magnitude decay alone is sufficient to write identity at the correct position.

(iii) **SPIRAL.** The descriptive analyses (Section 3.1) show that contextualized embeddings of the same phoneme become less aligned as their positions grow further apart, suggesting that the identity direction rotates with position. In addition, because DECAY converges to a baseline, it separates late positions only weakly, and a rotation provides a second positional signal that does not saturate. SPIRAL therefore applies a position-dependent rotation to the DECAY scaling, $\phi(p) = R(p)\, s(p)$, with $s(p)$ as in Eq. (4). We build $R(p)$ with the geometric construction of rotary position embeddings (Su et al., 2024), borrowing the construction but not its use as a transformer position encoding. The $2d$ dimensions are grouped into d pairs, and the plane spanned by each pair is rotated by an angle that grows linearly with position:

$$R(p) = \mathrm{diag}\left(R_1(p), \ldots, R_d(p)\right), \quad R_i(p) = \begin{pmatrix} \cos\theta_i & -\sin\theta_i \\ \sin\theta_i & \cos\theta_i \end{pmatrix}, \quad \theta_i = p\,\omega_i. \tag{5}$$

The only added parameters are the d frequencies ω_i, so different parts of the state can rotate at different rates. Identity is still written along a direction, but that direction turns linearly with position while its scale decays exponentially.

Unconstrained variants that fit the scaling at each position independently, and thus bound what the additive edit can achieve, are reported in Appendix E.

Subspace baseline. The three scaling functions assume that position is encoded in the magnitude and identity in the direction of a shared state. A common alternative is that each position occupies its own linear subspace of the state. SUBSPACE tests this alternative using a variant of distributed alignment search (DAS; Geiger et al., 2024), following Csordás et al. (2024). Instead of adding a vector, SUBSPACE overwrites part of the base state with content copied from a third sequence, the *source*, which already carries the target phoneme. It learns an orthogonal rotation of z into a basis in which the positional subspaces are axis-aligned, together with an assignment of each rotated dimension to a positional variable spanning k consecutive positions. The source agrees with the target only on the positions covered by the intervened variables and is resampled uniformly elsewhere, so it provides no context that the additive methods lack. We report $k \in \{1, 2\}$; the construction is given in Appendix D.

2.3 HUMAN INTRACRANIAL RECORDINGS

Data and preprocessing. We analyze high-gamma activity (70–150 Hz) recorded from 59 sound-responsive superior temporal gyrus (STG) electrodes, identified in 9 out of the 20 participants of the intracranial dataset of Lakretz et al. (2026), while they listened to 152 short spoken sentences (508 words, 1618 phonemes). Details are given in Appendix A.

Contextualized neural responses. Phoneme boundaries were obtained by forced alignment (McAuliffe et al., 2017) with manual correction. Let $\bar{r}(p) \in \mathbb{R}^{59}$ be the high-gamma amplitude of the selected electrodes, averaged over the interval of phoneme p. By analogy with Eq. (1), we define the contextualized neural response to the i-th phoneme of a word as

$$\Delta_i \;=\; \bar{r}(p_i) \;-\; \bar{r}(p_{i-1}). \tag{6}$$

This captures the contribution of each successive phoneme rather than the cumulative response. We excluded word-initial phonemes, for which Δ_i is undefined, and the first word of each sentence, which carries the sentence-onset response. This leaves 1110 responses. The positional prediction is tested by relating $\|\Delta_i\|$ to within-word position.

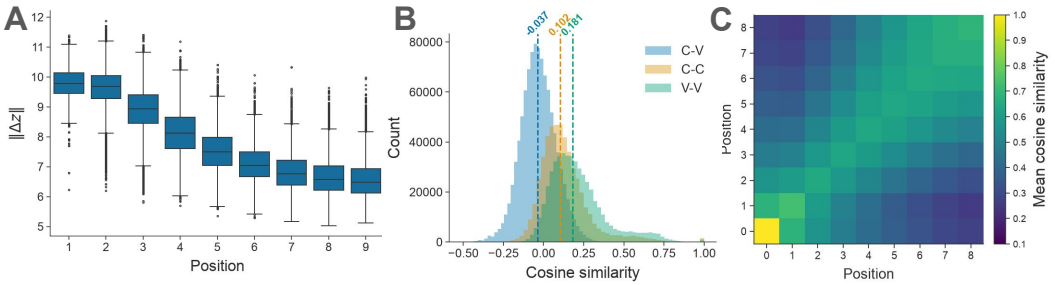


Figure 2: **Hypothesis-Free Analysis of the Geometry of Phoneme-Sequence Encoding.** *Phoneme-position encoding:* **(A)** Norm of the contextualized phoneme embedding Δz as a function of phoneme position. In accordance with the onion hypothesis, the norm is largest at the first position and decreases for later positions in the word. *Phoneme-identity encoding:* The onion hypothesis holds that identity is carried by the direction of Δz. **(B)** Distribution of cosine similarities between Δz vectors, split by within-class (C–C, V–V) and across-class (C–V) pairs, where C and V denote consonants and vowels; medians dashed. Within-class pairs are positively aligned, while across-class pairs are close to orthogonal, so direction carries identity information. **(C)** Mean cosine similarity between contextualized embeddings of the same phoneme at each pair of positions, averaged over the phoneme inventory. If direction were preserved across positions, as the onion hypothesis requires, all entries would be close to one. Instead, similarity varies systematically with position, indicating that the identity direction changes as position advances and motivating the spiral refinement.

Controlling for predictability. Later phonemes in a word are on average more predictable, and less predictable phonemes elicit larger cortical responses (Brodbeck et al., 2018; Donhauser & Baillet, 2020). A norm that declines with position could therefore reflect declining surprisal. To separate the two, we compute the cohort surprisal of each phoneme, $s(p_i) = -\log_2 P(p_i \mid p_1 \dots p_{i-1})$, over the entries of the CMU Pronouncing Dictionary (Weide, 1998) (details in Appendix A).

Identity analysis. For the identity prediction, we compute the cosine similarity between all pairs of responses and group the pairs into consonant–consonant, vowel–vowel, and consonant–vowel. We quantify the separation between within- and between-class similarities with the area under the ROC curve (AUC): the probability that a random within-class pair is more similar than a random between-class pair (0.5 is chance). As a rank statistic, AUC is insensitive to the scale and skew of the similarities and comparable across contrasts with different pair counts. Significance is assessed by permuting phoneme class labels (1000 permutations).

3 Results

3.1 The Onion Hypothesis Holds for Scale but Fails for Orientation

We first conducted a hypothesis-free descriptive analysis, examining the magnitude and direction of phoneme encodings in order to characterize the geometry of phoneme-sequence encoding empirically. The representations reproduced the broad onion pattern. The norm of the delta state was largest at the first position and decreased with position (Figure 2A), matching the magnitude decay the onion hypothesis predicts.

To test whether the organization of identity respects broader phonological structure, we grouped phonemes into the two major functional classes, consonants (C) and vowels (V). Direction did carry identity in the way the hypothesis proposes – within-type delta states (C-C or V-V pairs) were more similar to each other than cross-type ones (C-V pairs, Figure 2B).

Examining how identity is preserved across positions, however, revealed a departure from the onion account. If position were carried by magnitude alone, the contextualized embeddings of a given phoneme would point in the same direction at every position, and since cosine similarity ignores magnitude, their similarity would be close to one for any pair of positions. We therefore computed, for each phoneme, the mean cosine similarity between its embeddings at every pair of positions, and averaged over the phoneme inventory (Figure 2C). Similarity is high between nearby positions but

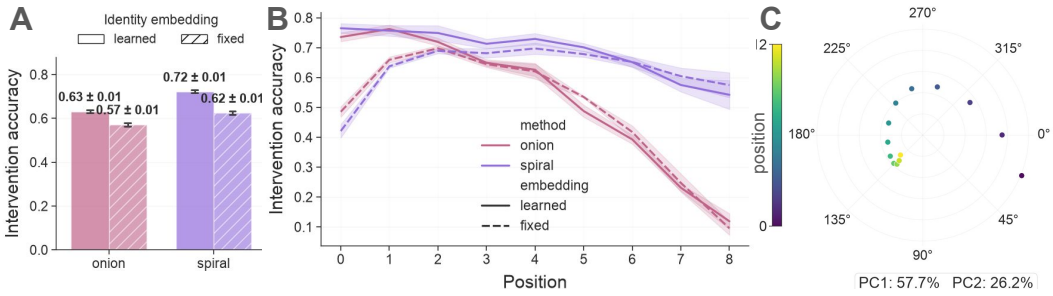


Figure 3: **The Spiral Modulation Outperforms the Onion Modulation. (A)** Mean intervention accuracy by method, with identity embeddings either learned or fixed. The spiral method outperforms the onion method in both settings **(B)** Intervention accuracy by position. The two methods are comparable over the first few positions, but the onion method degrades sharply from position 4 onward while the spiral method declines only gradually, indicating that the rotation matters most where magnitude decay alone has compressed the representation. **(C)** PCA of the learned modulation ϕ for the spiral method, in polar coordinates, one point per position (colour). The radius decays roughly exponentially with position, while the angle turns monotonically.

declines with the distance between them, approaching orthogonality for distant positions. The identity direction is therefore not fixed but changes systematically with position. This descriptive analysis alone cannot tell whether this change reflects a systematic rotation, as the spiral hypothesis proposes, or, for example, the contextual influence of the surrounding phonemes. The causal comparison in Section 3.2 is designed to separate the two.

Table 1: **Causal interventions show that the spiral model outperforms the onion and subspace models.** Intervention accuracy by method and target, mean ± standard deviation over four cross-validation seeds (Section 2.2).

	Subspace baseline		Scale based	
Target	DAS, $k = 1$	DAS, $k = 2$	Onion	Spiral
Pseudoword	0.000 ± 0.000	0.173 ± 0.016	0.631 ± 0.008	**0.723** ± 0.006
Real word	0.102 ± 0.005	0.227 ± 0.005	0.901 ± 0.004	**0.927** ± 0.003

3.2 Causal State Interventions Support Spiral Encoding

The descriptive analysis showed that position alters the direction of the contextualized embeddings as well as their magnitude. To test whether the model uses this rotation, we compared the positional scaling functions of Section 2.2 by editing the encoder state and scoring the frozen decoder on exact sequence match. We report separately the edits that turn a real word into a pseudoword and the edits in the opposite direction.

The spiral outperforms scale-only and subspace accounts. Spiral achieved the highest accuracy in both directions, above Onion and far above Subspace (Table 1). Figure 3A breaks down the comparison between Spiral and Onion for edits whose target is a pseudoword, with identity embeddings either fixed or fine-tuned (Section 2.2). Fine-tuning improves accuracy for both methods (Onion: 0.57 to 0.63; Spiral: 0.62 to 0.72), but Spiral outperforms Onion in both settings, so its advantage does not depend on the added flexibility of learned embeddings. Figure 3B shows that the advantage is concentrated at late positions, where Onion degrades sharply while Spiral declines only gradually. A decaying scale alone compresses late positions together, so an additional positional signal matters most there.

The advantage comes from the rotation, not from capacity. Adding the rotation to Decay raises accuracy from 0.64 to 0.72, whereas adding the same rotation to Onion lowers it from 0.65 to 0.62 (Appendix E.4). Since Spiral also has fewer parameters than Onion, its advantage cannot be attributed to additional capacity. The rotation helps only when combined with a monotonically decaying

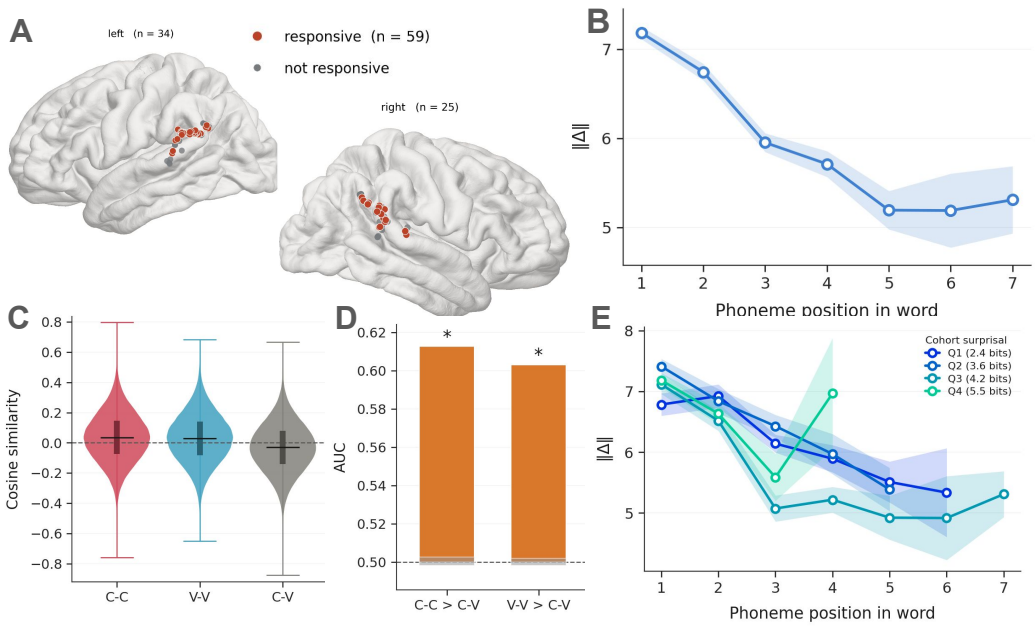


Figure 4: **Phoneme representation in the human superior temporal gyrus. (A)** STG electrodes on the MNI template. Red marks the 59 of 108 that respond to the speech stimuli and are used in all analyses. **(B)** Norm of phoneme representations against its within-word position, shading is SEM. **(C)** Cosine similarity between pairs of phoneme representations, split by consonant and vowel pair type. **(D)** AUC for the same contrasts, the probability that a within-class pair is more similar than a between-class pair (0.5 is chance, dashed). Grey band is the 95% interval of the label-permutation null (1000 shuffles), $*$ marks $p < 0.001$. **(E)** Same as (B) but recomputed separately within each of four surprisal quartiles, showing that the positional gradient is not carried by phoneme predictability.

scale, which is the combination the spiral hypothesis specifies. Moreover, the additive edit receives only the position and the identities of the base and target phonemes, not the surrounding phonemes, so the benefit of rotation cannot reflect contextual influence. Projecting the learned positional scaling of SPIRAL onto its leading principal components and reading it in polar coordinates shows the geometry directly: the radius decays with position while the angle turns steadily (Figure 3C).

Positions do not occupy separate subspaces. SUBSPACE fails almost completely at $k = 1$. This is expected if all positions share the same directions and are distinguished by magnitude and rotation, since such a code provides no position-specific subspace to swap. Csordás et al. (2024) report the same failure for their smallest recurrent models.

Intervention success depends on phonotactics. Finally, intervention success depends on the phonotactic legality of the word the edit creates. Accuracy rises from 0.722 with no filter to 0.866 once counterfactuals containing unattested phoneme trigrams are excluded (Appendix C). The positional code is therefore shaped by the statistics of the training language, an effect that random token sequences cannot reveal.

Together, these results show that the model uses the rotation suggested by the descriptive analysis to encode position. The positional code combines a decaying magnitude with a systematic rotation of the identity direction. It is not captured by magnitude alone or by separate positional subspaces, and it is shaped by the phonotactics of the training language.

3.3 HUMAN AUDITORY CORTEX ENCODES PHONEME POSITION IN RESPONSE MAGNITUDE

The model analyses yield concrete, quantitative predictions about how a neural population should bind phoneme identity to position. Testing them in the human auditory cortex requires recordings that resolve individual phonemes within a word, both in time and in space, which non-invasive methods cannot provide. Intracranial recordings from patients undergoing epilepsy monitoring offer

a rare opportunity to meet this requirement. We therefore asked whether the geometry found in the model is also present in the superior temporal gyrus (STG). We identified speech-responsive STG electrodes (Figure 4A) in our dataset (section 2.3) and applied the same construction as in the model, defining neural phoneme embeddings as the change in the high-gamma population response from one phoneme to the next (Eq. (6)). The onion and spiral accounts share two predictions: the magnitude of these embeddings should decrease with position, and their direction should carry phoneme identity.

Magnitude decreases with position. As in the model, the norm of the neural phoneme embeddings decreases monotonically with the phoneme's position in the word (Figure 4B; Spearman $\rho = -0.38$, $p = 1.6 \times 10^{-38}$, $n = 1110$). This is the signature predicted by both accounts. In natural speech, position might be confounded with predictability, because later phonemes in a word are on average more predictable, and predictable phonemes elicit weaker cortical responses. We therefore repeated the analysis within each of four cohort-surprisal quartiles. The decrease with position is present in every quartile (Figure 4E). The only departure is an increase at position 4 in the highest-surprisal quartile. The gradient is thus not a by-product of surprisal, indicating that human auditory cortex carries a magnitude code for within-word position.

Direction carries phoneme class. The second prediction is that the direction of the neural phoneme embeddings carries phoneme identity. Embeddings of phonemes from the same class are weakly but positively aligned, whereas pairs from different classes are close to orthogonal (Figure 4C). The AUC analysis further confirms that within-class pairs are more similar than across-class pairs (Figure 4D). Consonant–consonant (and also vowel–vowel) pairs are not reliably distinguished from each other (AUC 0.51), as expected if the two classes occupy comparably broad regions of the space.

An open prediction: rotation in cortex. In human STG, we thus find position-dependent magnitude decrease as in the onion and spiral accounts. The component that distinguishes the spiral account, a position-dependent rotation of each phoneme's direction, makes a further and more specific prediction. Testing this, however, requires many repetitions of the same phonemes at several within-word positions, which current existing data do not provide. Our study therefore provides an open prediction for future intracranial experiments.

4 Discussion

We asked whether the onion hypothesis, established on random token sequences, holds for a recurrent network trained on natural words, and whether its geometry has a counterpart in the human brain. In the model, position is carried by a decaying magnitude together with a systematic rotation, which Spiral captures more accurately than scale-only accounts while using fewer parameters. In human STG, the magnitude component is present and is not explained by phoneme predictability.

Our findings add to a growing body of work showing that neural language models encode linguistic structure in the geometry of their representations (Hewitt & Manning, 2019). A closely related polar geometry has been found for syntax (Diego-Simón et al., 2024; Diego-Simón et al., 2025), as well as for semantic structures (Diego-Simón et al., 2026). Our results can also be interpreted in terms of role–filler binding. In the additive intervention (Eq. (2)), the identity embedding is the filler, the positional scaling $\phi(p)$ is the role vector, and elementwise multiplication binds them. This retains only the diagonal of their tensor product (Smolensky, 1990), so the onion and spiral accounts are both compressed tensor product representations (TPRs) that differ only in their role vectors. Phoneme–position binding is thus another case in which neural networks spontaneously learn TPRs (McCoy et al., 2018; Soulos et al., 2020; McCoy et al., 2026). Together with the polar code for syntax and semantics, this suggests a general scheme in which distinct geometric components encode distinct variables of structured representations.

The geometry also links to serial-order memory. The onion account is a primacy gradient (Hurlstone et al., 2014), and the rotation adds explicit position marking, as in oscillator models (Brown et al., 2000). It also explains the serial-position curve: the first phoneme is written at the largest scale, making it robust to interference from later phonemes (primacy), while the last phoneme, though written at the smallest scale, undergoes no later interference (recency).

Together, these findings suggest that neural systems can solve the binding problem for sequences without dedicated positional slots, by binding each filler to its role through the magnitude and direction of a shared state.

ETHICS STATEMENT

The intracranial recordings we analyze were collected for clinical purposes in patients undergoing pre-surgical monitoring for epilepsy, under the protocols and informed consent procedures reported in the original data descriptor (Lakretz et al., 2026). We performed no new recordings and analyzed only de-identified data. No other aspect of this work raises concerns under the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

The recurrent model, the delta-state analyses, the intervention operator and all modulation parameterizations are described in Section 2, with the remaining parameterizations, the optimization settings and the interchange-intervention construction in Appendices B, D and E. The intracranial preprocessing, electrode selection, delta-phoneme definition, cohort-surprisal model and permutation tests are given in Section 2.3. Code for the model analyses and interventions is provided as anonymized supplementary material and will be released publicly upon acceptance.

AI USE STATEMENT

Generative AI tools were used to assist with editing and formatting of the manuscript text, to improve the code pipeline for the analyses, and to refine the visual styling of the figures. They were not used to generate research ideas, to design experiments, or to interpret results. We have reviewed all AI-assisted work, and all code produced with such assistance was verified against the analyses it implements. We take responsibility for the final content of this work, including text, claims and artifacts.

REFERENCES

Timothy E.J. Behrens, Timothy H. Muller, James C.R. Whittington, Shirley Mark, Alon B. Baram, Kimberly L. Stachenfeld, and Zeb Kurth-Nelson. What is a cognitive map? organizing knowledge for flexible behavior. *Neuron*, 100(2):490–509, 2018. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2018.10.002. URL https://www.sciencedirect.com/science/article/pii/S0896627318308560.

Christian Brodbeck, L. Elliot Hong, and Jonathan Z. Simon. Rapid transformation from auditory to linguistic representations of continuous speech. *Current Biology*, 28(24):3976–3983.e5, 2018. ISSN 0960-9822. doi: 10.1016/j.cub.2018.10.042.

Gordon Brown, Tim Preece, and Charles Hulme. Oscillator-based memory for serial order. *Psychological Review*, 107:127–181, 01 2000. doi: 10.1037/0033-295X.107.1.127.

Declan Iain Campbell, Sunayana Rane, Tyler Giallanza, C. Nicolò De Sabbata, Kia Ghods, Amogh Joshi, Alexander Ku, Steven M Frankland, Thomas L. Griffiths, Jonathan D. Cohen, and Taylor Whittington Webb. Understanding the limits of vision language models through the lens of the binding problem. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=Q5RYn6jagC.

Francois Chollet, Mike Knoop, Gregory Kamradt, Bryan Landers, and Henry Pinkard. Arc-agi-2: A new challenge for frontier ai reasoning systems, 2026. URL https://arxiv.org/abs/2505.11831.

G. N. Clements. *The role of the sonority cycle in core syllabification*, pp. 283–333. Papers in Laboratory Phonology. Cambridge University Press, 1990.

Róbert Csordás, Christopher Potts, Christopher D Manning, and Atticus Geiger. Recurrent neural networks learn to store and generate sequences using non-linear representations. In Yonatan Belinkov, Najoung Kim, Jaap Jumelet, Hosein Mohebbi, Aaron Mueller, and Hanjie Chen (eds.), *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 248–262, Miami, Florida, US, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.blackboxnlp-1.17. URL https://aclanthology.org/2024.blackboxnlp-1.17/.

Daniel Dager, Robin Sobczyk, Emmanuel Chemla, and Yair Lakretz. A neural model for word repetition, 2025. URL https://arxiv.org/abs/2506.13450.

Pablo Diego-Simón, Stéphane D' Ascoli, Emmanuel Chemla, Yair Lakretz, and Jean-Rémi King. A polar coordinate system represents syntax in large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 105375–105396. Curran Associates, Inc., 2024. doi: 10.52202/079017-3344. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/be36e50757bf9cd280aa74f89a7d1c23-Paper-Conference.pdf.

Pablo J Diego-Simón, Emmanuel Chemla, Jean-Rémi King, and Yair Lakretz. Probing syntax in large language models: Successes and remaining challenges. 2025.

Pablo J Diego-Simón, Pierre Orhan, Emmanuel Chemla, Yair Lakretz, and Jean-Rémi King. Polar probe linearly decodes semantic structures from llms. 2026.

Peter W. Donhauser and Sylvain Baillet. Two distinct neural timescales for predictive speech processing. *Neuron*, 105(2):385–393.e9, 2020. doi: 10.1016/j.neuron.2019.10.019.

Jeffrey L. Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990. ISSN 0364-0213. doi: https://doi.org/10.1016/0364-0213(90)90002-E. URL https://www.sciencedirect.com/science/article/pii/036402139090002E.

Jiahai Feng and Jacob Steinhardt. How do language models bind entities in context? In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.17191.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 9574–9586. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/4f5c422f4d49a5a807eda27434231040-Paper.pdf.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In Francesco Locatello and Vanessa Didelez (eds.), *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 160–187. PMLR, 01–03 Apr 2024. URL https://proceedings.mlr.press/v236/geiger24a.html.

Klaus Greff, Sjoerd van Steenkiste, and Jürgen Schmidhuber. On the binding problem in artificial neural networks, 2020. URL https://arxiv.org/abs/2012.05208.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling (COLM)*, 2024. URL https://arxiv.org/abs/2312.00752.

Albert Gu, Karan Goel, and Christopher Re. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=uYLFoz1vlAC.

Brian C. Hall. *Lie Groups, Lie Algebras, and Representations: An Elementary Introduction*, volume 222 of *Graduate Texts in Mathematics*. Springer, 2003.

John Hewitt and Christopher D Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138, 2019.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9: 1735–1780, 11 1997. doi: 10.1162/neco.1997.9.8.1735.

J J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982. doi: 10.1073/pnas.79.8.2554. URL https://www.pnas.org/doi/abs/10.1073/pnas.79.8.2554.

Mark J. Hurlstone, Graham J. Hitch, and Alan D. Baddeley. Memory for serial order across domains: An overview of the literature and directions for future research. *Psychological Bulletin*, 140(2): 339–373, 2014. ISSN 1939-1455. doi: 10.1037/a0034221.

Mete Ismayilzada, Defne Circi, Jonne Sälevä, Hale Sirin, Abdullatif Köksal, Bhuwan Dhingra, Antoine Bosselut, Duygu Ataman, and Lonneke van der Plas. Evaluating morphological compositional generalization in large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1270–1305, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.59. URL https://aclanthology.org/2025.naacl-long.59/.

Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2873–2882. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/lake18a.html.

Yair Lakretz, Naama Friedmann, Jean-Rémi King, Emily Mankin, Anthony Rangel, Ariel Tankus, Stanislas Dehaene, and Itzhak Fried. Modality-specific and amodal language processing by single neurons. *eLife*, June 2026. doi: 10.7554/eLife.110689.1. URL https://doi.org/10.7554/eLife.110689.1. Reviewed preprint.

Matthew K. Leonard, Kristofer E. Bouchard, Claire Tang, and Edward F. Chang. Dynamic encoding of speech sequence probability in human temporal cortex. *Journal of Neuroscience*, 35(18): 7203–7214, 2015. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.4100-14.2015. URL https://www.jneurosci.org/content/35/18/7203.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Proceedings of Interspeech 2017*, pp. 498–502, 2017. doi: 10.21437/Interspeech.2017-1386.

R Thomas McCoy, Tal Linzen, Ewan Dunbar, and Paul Smolensky. Rnns implicitly implement tensor product representations. 2018.

R Thomas McCoy, Paul Soulos, Tal Linzen, and Paul Smolensky. The emergent symbolic structure of artificial neural networks. *arXiv preprint arXiv:2608.29530*, 2026.

N. Mesgarani, C. Cheung, K. Johnson, and E. F. Chang. Phonetic feature encoding in human superior temporal gyrus. *Science*, 343(6174):1006–1010, Jan 2014a. doi: https://doi.org/10.1126/science.1245994.

Nima Mesgarani, Connie Cheung, Keith Johnson, and Edward F. Chang. Phonetic Feature Encoding in Human Superior Temporal Gyrus. *Science*, 343(6174), February 2014b. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1245994. URL https://www.science.org/doi/10.1126/science.1245994.

Josef Parvizi and Sabine Kastner. Promises and limitations of human intracranial electroencephalography. *Nature Neuroscience*, 21, 03 2018. doi: 10.1038/s41593-018-0108-2.

Steven Piantadosi. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 21:1112–1130, 03 2014. doi: 10.3758/s13423-014-0585-6.

Victor Rambaud, Salvador Mascarenhas, and Yair Lakretz. Mapformer: Self-supervised learning of cognitive maps with input-dependent positional embeddings, 2026. URL https://arxiv.org/abs/2511.19279.

Adina L. Roskies. The binding problem. *Neuron*, 24(1):7–9, 1999. ISSN 0896-6273. doi: https://doi.org/10.1016/S0896-6273(00)80817-X. URL https://www.sciencedirect.com/science/article/pii/S089662730080817X.

Paul Smolensky. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46(1):159–216, 1990. ISSN 0004-3702. doi: https://doi.org/10.1016/0004-3702(90)90007-M. URL https://www.sciencedirect.com/science/article/pii/000437029090007M.

Paul Soulos, R Thomas McCoy, Tal Linzen, and Paul Smolensky. Discovering the compositional structure of vector representations with role learning networks. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 238–254, 2020.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with Rotary Position Embedding. *Neurocomputing*, 568:127063, February 2024. ISSN 09252312. doi: 10.1016/j.neucom.2023.127063. URL https://linkinghub.elsevier.com/retrieve/pii/S0925231223011864.

Michael S. Vitevitch and Paul A. Luce. Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language*, 40(3):374–408, 1999. ISSN 0749-596X. doi: https://doi.org/10.1006/jmla.1998.2618. URL https://www.sciencedirect.com/science/article/pii/S0749596X98926183.

Robert L. Weide. The CMU Pronouncing Dictionary. Carnegie Mellon University, 1998. URL http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1819–1862, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.102. URL https://aclanthology.org/2024.naacl-long.102/.

Zhengxuan Wu, Atticus Geiger, Thomas Icard, Christopher Potts, and Noah D. Goodman. Interpretability at scale: Identifying causal mechanisms in alpaca. In *Advances in Neural Information Processing Systems*, volume 36, pp. 78205–78226, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f6a8b109d4d4fd64c75e94aaf85d9697-Abstract-Conference.html.

Han Gyol Yi, Matthew K. Leonard, and Edward F. Chang. The encoding of speech sounds in the superior temporal gyrus. *Neuron*, 102(6):1096–1110, 2019. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2019.04.023. URL https://www.sciencedirect.com/science/article/pii/S0896627319303800.

LIMITATIONS

Several extensions would strengthen our conclusions. On the modeling side, testing other architectures, hidden sizes and languages would establish how general the modulation is, and whether the per-position rotation grows with capacity. Recordings with higher signal-to-noise ratio and many repetitions of the same phonemes at controlled positions would allow a direct test of the rotational component, which distinguishes the spiral from the onion account. Such data would also support finer identity contrasts than consonant–vowel separability, which is well established in STG (Mesgarani et al., 2014b). That the predicted signatures already emerge in a modest intracranial dataset (1110 phonemes from 152 sentences) makes these tests promising.

A Intracranial Recording Details

Dataset. We analyze the auditory sentence runs of the intracranial dataset of Lakretz et al. (2026), in which 20 participants implanted with depth electrodes for epilepsy monitoring listened to short spoken sentences. The analyzed set comprises 152 sentences, 508 words, and 1618 phonemes. Intracranial electroencephalography resolves cortical activity at millisecond and millimetre scales simultaneously and with a high signal-to-noise ratio, a combination that scalp recordings and functional imaging cannot achieve (Parvizi & Kastner, 2018). Its main limitations are the clinical population and the sparse, clinically driven electrode coverage.

Recording and preprocessing. For each participant, recordings were notch filtered at 60 Hz, clipped at $5\times$IQR, re-referenced to the median of the channel set and then within each electrode shaft, and screened automatically for bad channels. High-gamma activity, a proxy for local population firing, was computed as the analytic amplitude of the 70–150 Hz band using the Hilbert transform. Epochs were cut from -0.5 to 1.5 s around sentence onset and z-scored against the $[-0.5, 0]$ s baseline.

Electrode selection. Contacts were assigned to a region by their MNI coordinate under the AAL3 atlas, taking Temporal_Sup in both hemispheres. Responsiveness was tested with a one-sample cluster permutation test of (response $-$ baseline) > 0, with baseline $[-0.3, -0.01]$ s, response $[0, 1]$ s, a cluster-forming threshold of $p < 0.05$ one-tailed, 1000 sign flips and a cluster threshold of $p < 0.01$. Each channel's delay is the onset of its earliest significant cluster. Of the 108 superior temporal gyrus (STG) electrodes, we retain the 59 that respond significantly to sentence onset (cluster permutation test against baseline). These come from 9 of the 20 participants. Each electrode's signal is shifted by its own response latency (median 43 ms).

Cohort surprisal. The lexicon is the CMU Pronouncing Dictionary (Weide, 1998) augmented with every pronunciation actually presented, so that no observed prefix is unseen, with add-0.5 smoothing applied to the cohort continuation counts.

B Intervention Training and Dataset Details

The encoder and decoder each use one LSTM layer with hidden size $d = 128$, so the concatenated encoder state $z = [h; c]$ has $2d = 256$ dimensions.

Settings. In addition to the two choices of target reported in the main text, counterfactual and word, we test a third setting in which the source sequence the state is taken from and the target sequence are both real English words differing at a single phoneme (Table 1).

Optimization. Interventions are trained with Adam, minimizing the cross-entropy between the target sequence and the decoder output produced from the edited state. Training uses early stopping on validation accuracy, halting after about 25 epochs on average. Unless stated otherwise, appendix comparisons are point estimates from a single trained model, with the dataset fixed across runs by a shared seed, whereas Table **??** in the main text reports means over cross-validation seeds.

Embeddings. The phoneme embeddings that seed the difference $(x_{\text{new}} - x_{\text{old}})$ in Equation 2 are squashed by a $\tanh$ before differencing. When the embedding is initialized from the repetition model, whose embedding dimension is $d = 128$, it is duplicated by concatenation with itself to match the $2d = 256$-dimensional state, which adds no trainable parameters. Held fixed, the mean-Δz and base-model embeddings are indistinguishable, and training either jointly with ϕ raises accuracy well above both (Table 2). Which initialization is best once trained is not stable across parameterizations, so we draw nothing from that ordering and use the mean-Δz initialization throughout for comparability.

C Phonotactic Legality and Intervention Success

Counterfactual pairs are built by replacing a phoneme at random, so many of the resulting sequences are ones no English word could contain. We define two filters over them. The consonant-vowel

Table 2: Effect of phoneme embedding initialization and joint training, for the spiral method on state z with the counterfactual as target.

Embedding init.	Joint train	Accuracy
mean Δz	✓	0.720 ± 0.011
	✗	0.623 ± 0.010
base-model emb.	✓	$\mathbf{0.807} \pm 0.009$
	✗	0.628 ± 0.008
from scratch	✓	0.750 ± 0.009

Table 3: Intervention accuracy for the main methods in the real-to-real setting, where both source and target sequence are real words from train set. Embeddings are initialized from the mean Δz and trained jointly. The unconstrained method is the low-rank form at rank 8.

Method	Accuracy
Unconstrained	0.995
Spiral	0.996
Onion	0.994
DAS	0.568
DAS bigram	0.704

filter discards substitutions that violate the consonant-vowel patterns seen in training, and the n-gram filter discards those creating phoneme bigrams or trigrams absent from real-word training sequences. Crossing the two gives the design reported here.

Artificial token sequences are unconstrained, so models evaluated on synthetic data never encounter phonological rules. In natural language, substituting a phoneme at random frequently generates strings that violate the constraints of the language. We therefore asked whether intervention success depends not only on the geometry of the state vector but also on whether the edited state lands in a phonotactically valid region of the decoder's representation space. We filtered counterfactual targets by structural validity (Table 4). With no filter, intervention accuracy was 0.722. Enforcing consonant-vowel structure alone raised it to 0.743, whereas rejecting unattested n-grams had a considerably larger effect, reaching 0.772 for bigrams and 0.866 for trigrams. The n-gram filters therefore contributed substantially more than the local CV constraint, which shows that the structure governing what the decoder can read out extends beyond immediately neighbouring phonemes. The two constraints were also largely redundant, since CV added to the bigram filter (0.772 to 0.790) but nothing to the trigram filter (0.866 in both cases), as expected if trigram statistics already capture the CV alternation. Because random sequences lack these higher-order constraints, this dependence cannot appear in the synthetic setting where the onion hypothesis was originally derived.

Table 4: Effect of each phonotactic filter and their combinations, for the spiral model with trained mean-Δz embeddings. CV enforces the consonant-vowel patterns of real words and N-gram rejects unattested bigrams (2) or trigrams (3).

CV	N-gram	Accuracy
✗	✗	0.722 ± 0.009
✓	✗	0.743 ± 0.011
✗	2	0.772 ± 0.005
✓	2	0.790 ± 0.006
✗	3	$\mathbf{0.866 \pm 0.005}$
✓	3	0.866 ± 0.010

D Interchange Intervention Details

The interchange baseline of Section 2 follows Csordás et al. (2024), a variant of distributed alignment search (Geiger et al., 2024; Wu et al., 2023). Each example supplies a base sequence b, whose state is edited, and a source sequence s, whose state is copied from.

The hypothesis under test is that each position occupies a linear subspace of the encoder state. The subspaces need not be axis-aligned, so the method learns an orthogonal $\mathbf{R} \in \mathbb{R}^{2d \times 2d}$ that rotates z into a basis where they are, together with an assignment matrix $\mathbf{A} \in \{0,1\}^{2d \times (V+1)}$ that gives each rotated dimension to one of V positional variables or to a final column of dimensions that are

Table 5: Data for replacing the phoneme at position $p = 3$ of *plant* with /IH/. Bold marks the edit, underlining marks the window on which the source must agree with the target, which widens with k. Every other source position is resampled. The additive method needs no source sequence.

Method	Base	Source	Target
Additive	P L **AE** N T	—	P L **IH** N T
DAS, $k = 1$	P L **AE** N T	K R **IH** S P	P L **IH** N T
DAS, $k = 2$	P L **AE** N T	F L **IH** N CH	P L **IH** N T

never intervened on. Subspace sizes are therefore learned rather than fixed. $\mathbf{A}$ is the row-wise hard Gumbel-softmax of a real-valued $\hat{\mathbf{A}}$ with a straight-through estimator, penalized toward the final column.

Let S be the variables the edit touches and $\mathbf{m} = \sum_{v \in S} \mathbf{A}[\cdot, v]$ the dimensions they own. The intervention replaces those dimensions with the source's and rotates back,

$$\begin{aligned} \hat{\tilde{z}} &= \mathbf{m} \odot \mathbf{R} z^{\mathrm{s}} + (\mathbf{1} - \mathbf{m}) \odot \mathbf{R} z^{\mathrm{b}}, \\ \hat{z} &= \mathbf{R}^{\top} \hat{\tilde{z}}. \end{aligned} \tag{7}$$

$\mathbf{R}$ and $\hat{\mathbf{A}}$ are trained through the frozen decoder in the same way as ϕ. A variable spans k consecutive positions, giving $V = P - k + 1$ of them for a maximum sequence length P, and S holds every variable whose window covers the edited position. Unlike Equation 2, which sees only the position and the two phonemes, Equation 7 sees a full encoder state, so the source has to be built to leak no context, as described in Section 2.2.

E Additional Modulation Parameterizations

Beyond the onion and spiral methods of Section 2, we evaluate the parameterizations below. Throughout, $m = 2d = 256$ is the state dimension and p is the position of the edit.

E.1 Unconstrained Methods

Unconstrained parameterizations store $\phi(p)$ independently for each position and impose no relationship between positions. They give the most flexible fit and serve two roles, an upper bound on what the additive edit can achieve, and a diagnostic, since inspecting the learned vectors shows whether the position transformation follows a regular pattern.

Per-position. Each position holds its own vector,

$$\phi(p) = \Phi_p, \qquad \Phi \in \mathbb{R}^{P \times 2d}, \tag{8}$$

where P is the maximum sequence length. This is the most flexible form and uses $P \times 2d$ parameters.

Low-rank. To keep positions independent while reducing parameters, the modulation matrix is factored,

$$\phi(p) = C_p B^{\top}, \qquad C \in \mathbb{R}^{P \times r}, \quad B \in \mathbb{R}^{2d \times r}, \tag{9}$$

with $r \ll 2d$. The rank r trades flexibility against parameter count, and sharing B across positions also regularizes the fit.

E.2 Scale-Only Methods

Linear. The onion form without its exponential term,

$$\phi(p) = \beta\, p + b, \qquad \beta, b \in \mathbb{R}^m. \tag{10}$$

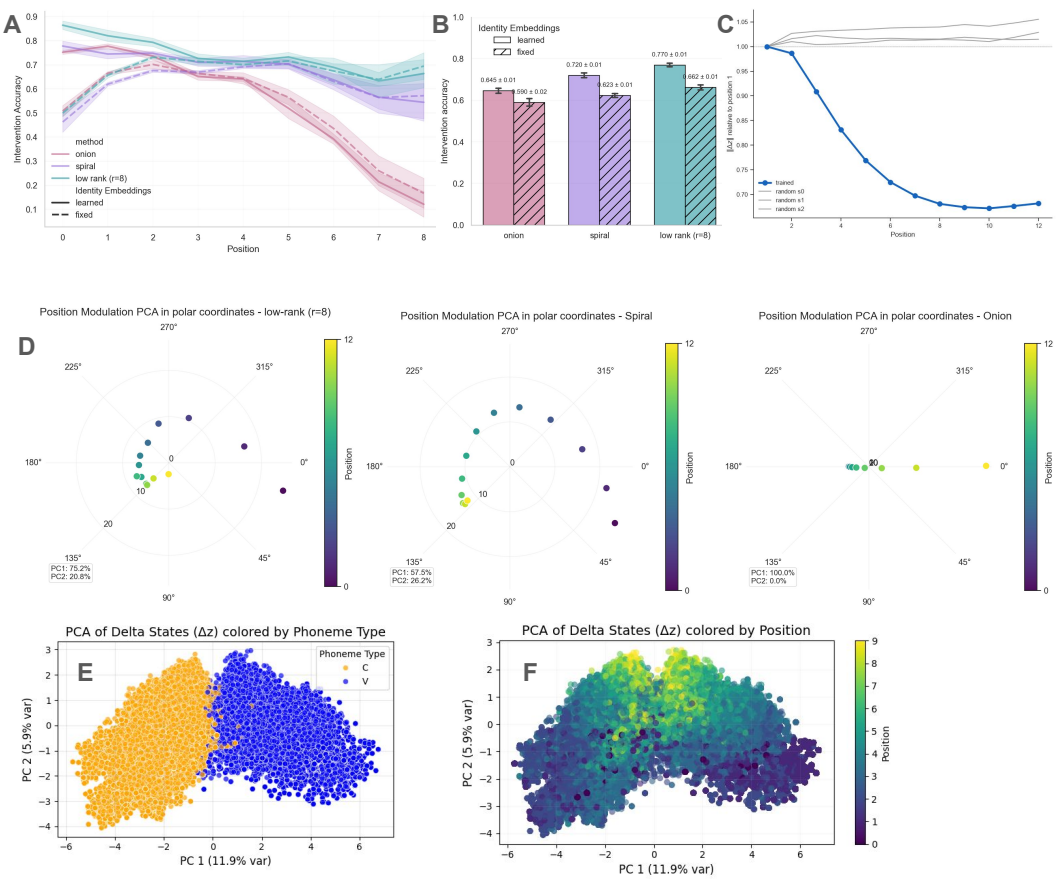


Figure 5: **(A, B)** Intervention accuracy by position for each method, including the unconstrained low-rank-8 method. The spiral method is competitive with the unconstrained method using far fewer parameters. **(C)** Mean $\|\Delta z\|$ by position for the trained repeat model and three randomly initialised encoders of the same architecture, each curve normalised to its own position-1 value (untrained weights sit at an arbitrary scale, so only the shape is comparable). The trained profile decays; the untrained ones are flat. The decay in Figure 2A is learned, is not a generic consequence of LSTM dynamics, but a learned modulation that carries position. **(D)** Learned modulation vectors ϕ for the main methods in PCA space. The rotation is present in the unconstrained and spiral methods and absent from the onion method. **(E)** Consonants and vowels separate in PCA space. **(F)** Earlier positions are more separated while later positions converge toward the center, which explains why fixed embeddings underperform at early positions.

E.3 Scale-and-Rotation Methods

Lie-group spiral. A single matrix exponential rotates all coordinates together (Hall, 2003),

$$R(p) = \exp(p\,A), \qquad A = W - W^\top, \quad A \in \mathbb{R}^{m\times m}, \tag{11}$$

with $\phi(p) = R(p)\,s(p)$ and s the Decay scale. Skew-symmetry of A makes $R(p)$ orthogonal at every position. This form can couple any pair of coordinates rather than the fixed planes of the spiral method, so it is strictly more expressive, at the cost of $m(m-1)/2$ parameters and a matrix exponential per position.

E.4 Isolating the Contribution of the Rotation

Crossing the scale parameterization with the presence of a rotation separates the contribution of the rotation from that of the scale (Table 6). Adding the rotary construction to the exponentially decaying scale raises accuracy from 0.64 to 0.72, whereas adding the same rotation to the onion scale

lowers it from 0.65 to 0.62. The rotation therefore helps only when it is applied to a scale that decays monotonically, which is what the spiral hypothesis describes. The onion scale carries a linear term in addition to its exponential one, so its direction already drifts with position, and imposing a further rotation on it appears to overparameterize the transformation rather than add to it.

Table 6: Intervention accuracy for each combination of scale parameterization and rotation, with the counterfactual as target. The rotation improves the monotonically decaying scale and degrades the onion scale.

	Rotation	
Scale	None	Rotary
Onion	0.65	0.62
Exponential decay	0.64	**0.72**

E.5 Comparison of Parameterizations

Table 7 compares all parameterizations on the source-to-counterfactual set. Low-rank at rank 4 and 8 is best, slightly above the unconstrained per-position look-up, which is consistent with regularization, since the full look-up has one free vector per position and can overfit the training counterfactuals while a low rank shares parameters across positions and generalizes better. Among the functional methods the spiral form is strongest, above the Lie-group spiral and above the scale-only onion and decay forms, and decay does not improve on onion. On the fully filtered set the simplest baselines, linear and rank 1, remain close to the rest, which reflects how much the filters ease the task.

Table 7: All parameterizations on the source-to-counterfactual set, with trained and fixed mean-Δz embeddings, under the repeat-only filter and under all filters.

	Trained emb.		Fixed emb.	
Method	Repeat	All filt.	Repeat	All filt.
Low-rank-8	**0.780**	**0.898**	0.678	0.807
Per-position	0.773	0.878	0.679	0.800
Low-rank-4	0.736	0.889	0.663	0.802
Spiral	0.729	0.866	0.632	0.770
Low-rank-2	0.723	0.865	0.660	0.788
Lie spiral	0.721	0.854	0.601	0.747
Linear	0.658	0.794	0.590	0.739
Decay	0.651	0.807	0.599	0.707
Onion	0.650	0.815	0.601	0.769
Low-rank-1	0.635	0.768	0.554	0.715

F Which State Carries the Code

The concatenated state $z = [h; c]$ and the cell state c alone give nearly identical accuracy, while editing the hidden state h alone fails completely (Table 8). The positional code the intervention exploits therefore lives in the cell state, which is consistent with the cell state carrying the memory the decoder reads.

Table 8: Intervention accuracy by the state edited, on the source-to-counterfactual set with all filters, the rank-8 method, across embedding settings. When the intervention edits c, the decoder receives the source sequence's h together with the intervened c', and the reverse holds for editing h.

	mean Δz		repetition emb.	
State	Trained	Fixed	Fixed	Trained
$z = [h; c]$	0.899	0.807	0.739	0.893
c (cell)	0.888	0.768	0.737	0.896
h (hidden)	0.000	0.000	0.000	0.000