BOUNDED INTEGRATION WINDOWS GATE HYBRID CONTINUOUS–SPARSE READOUTS IN A SPIKE-CONVERTED TRANSFORMER ALIGNED WITH LANGUAGE CORTEX

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models and the human brain both turn streams of words into structured meaning. A growing body of work shows that the internal activations of Transformer language models can linearly predict brain responses recorded while people listen to natural speech. How far this convergence reflects shared computation between language models and human brain remains open. Two basic properties of language cortex have been hard to test against models. Cortical regions integrate information over bounded, region-specific timescales rather than indefinitely. Cortical activity also contains sparse, event-like structure, which current model–brain comparisons impose by post-hoc analysis rather than read out from the model. The obstacle is instrumental: standard Transformers expose only continuous activations and a fixed context window. We remove it with a same-weight artificial neural network (ANN) $\leftrightarrow$ spiking neural network (SNN) substrate. A pretrained Transformer is converted into a spiking network without retraining, keeping the original weights but unfolding computation over discrete timesteps. Continuous, sparse, and hybrid representations can then be read from a single forward pass, and the model-side integration window becomes a tunable scalar. Applied to naturalistic fMRI from 49 listeners, the substrate reveals three coupled phenomena. Brain alignment saturates inside a bounded integration plateau, where the model also shifts toward more compositional layers, a profile that replicates on a second backbone and conversion family. Within this plateau, the hybrid continuous–sparse readout predicts held-out responses better than a matched random control, while a redundant signed variant does not. The size of this advantage is non-monotonic in unfolding time and peaks earlier under faster leak across all twelve language parcels, while the relative hybrid lift descriptively tracks the cortical hierarchy of temporal receptive windows. Together, these results are consistent with a hybrid continuous–sparse code that is useful only inside a finite integration window, and they suggest that bounded temporal dynamics—long studied in spiking systems for biological reasons—may also act as a useful inductive bias for representations that align with the brain.

1 INTRODUCTION

Large language models and the human language network both integrate words into context-dependent meaning, and continuous Transformer hidden states linearly predict cortical responses to naturalistic language (Huth et al., 2016; Schrimpf et al., 2021; Caucheteux & King, 2022; Goldstein et al., 2022). The convergence is reproducible across datasets, languages, and recording modalities, yet it has been characterised along one axis: how well a continuous representation with a fixed architectural window predicts BOLD. Two long-standing observations about language cortex sit outside this axis. Cortical regions integrate information over bounded, region-specific timescales arranged along a temporal hierarchy that lengthens from auditory toward higher-order regions (Hasson et al., 2008; Lerner et al., 2011; Honey et al., 2012). Cortical activity also contains sparse, event-like structure, with step-like state transitions at narrative event boundaries (Baldassano et al., 2017). Neither has a counterpart in

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.

the encoding pipeline, where the model produces one continuous vector per token and the integration window is fixed by the architecture.

The obstacle is instrumental. Comparing continuous and sparse representations across two models confounds format with weights, training data, and architecture. Standard Transformers expose no sparse readout, and changing context length changes the input. Asking whether language cortex is better predicted by a continuous state, by sparse event-like recruitment, or by a hybrid, and whether the answer depends on the integration timescale, therefore needs a substrate in which both representations come from one forward pass with shared weights and the integration window is a controllable scalar.

Such a substrate exists in work aimed at energy-efficient inference. ANN-to-SNN conversion replaces the activations of a pretrained network with spikes unfolded over discrete timesteps (Rueckauer et al., 2017). SpikedAttention (Hwang et al., 2024) converts a matrix-attention BERT (Ming et al., 2023) without training, giving an ANN↔SNN pair with bit-identical linear weights whose dynamics are set by an unfolding length T and a leak base τ. The forward pass yields a continuous trace (the leak-weighted spike sum) and a sparse recruitment axis (the firing count), so continuous, sparse, and hybrid readouts are views of the same computation. T and τ are model-side parameters: the text input is fixed, and they change only how the converted network unfolds it. Sparse readouts also need a control that continuous features have not needed, because a matched random-discrete block can raise squared-correlation brain score without carrying stimulus information (Feghhi et al., 2024); we therefore test every hybrid claim on held-out variance against such a permutation control.

Applied to naturalistic fMRI from 49 listeners, the probe reveals three coupled phenomena. First, brain alignment from the continuous trace rises with T and saturates inside a bounded $T \times \tau$ plateau at the level of the parent ANN, where the preferred encoder layer shifts from shallow to deep; the profile holds in every listener and replicates on a standard BERT converted with an independent conversion family (Chen et al., 2026). Second, inside the plateau the hybrid continuous–sparse readout predicts held-out responses better than the matched random control on 13/13 language-network ROIs, whereas a signed-spike variant does not: the trace already carries signed firing, which leaves unsigned recruitment as the readout that a linear encoder cannot rebuild from it. Third, the size of this advantage is non-monotonic in T and peaks earlier under faster leak, in the language network and in all twelve language parcels, as a finite model-side integration window would predict. The relative hybrid lift is largest in posterior-temporal and middle-frontal parcels, a pattern descriptively consistent with the cortical hierarchy of temporal receptive windows.

Read together, the three phenomena are consistent with a hybrid continuous–sparse code that is useful only inside a finite integration window. A spike-converted Transformer reaches the alignment of its parent only inside its bounded $T \times \tau$ plateau, so the bounded dynamics that motivate spiking systems on biological grounds coincide here with the regime in which the model is informative about the brain.

2 Related work

Voxelwise encoding with continuous Transformers. Voxelwise encoding models link language-model representations to brain activity through a regularised linear map evaluated on held-out data (Naselaris et al., 2011). After early semantic atlases from word-level features (Huth et al., 2016), Transformer hidden states reached near-ceiling predictivity on language-network benchmarks (Schrimpf et al., 2021) and predicted ECoG and fMRI responses to natural speech (Goldstein et al., 2022; Caucheteux & King, 2022), with layer-depth and training effects on the corpus we use here (Pasquiou et al., 2022). The model-side feature space in this literature is real-valued, with discrete structure imposed downstream by clustering or thresholding. Brain scores can also be raised by features that carry little linguistic content (Feghhi et al., 2024), which motivates our control.

Continuous and discrete coding in human language cortex. The event-segmentation literature reports discrete cortical transitions during naturalistic comprehension, with regions arranged along a hierarchy of temporal receptive windows that lengthens from early auditory cortex toward higher-order temporal, parietal, and frontal areas (Hasson et al., 2008; Lerner et al., 2011; Honey et al., 2012; Baldassano et al., 2017). Continuous and discrete views have seldom been tested inside one feature space whose components come from the same forward pass and weights.

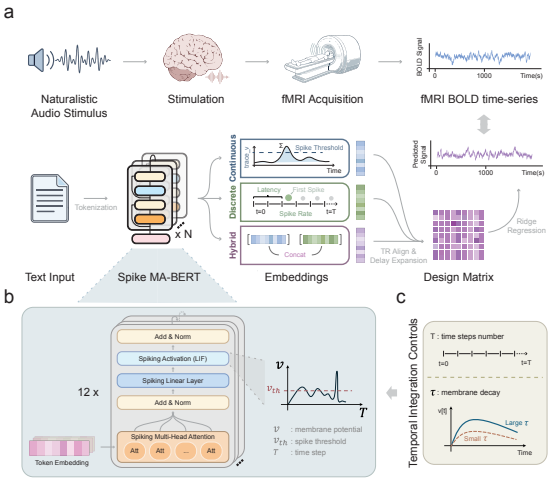

Figure 1: **Controlled SpikeNN encoding framework with modular representation and temporal controls. (a)** Naturalistic stimuli are processed by fMRI and a Spiked-MA-BERT model in parallel. From a single forward pass, we extract continuous (`trace_v`), discrete (e.g., `spike_rate`, `latency`), and hybrid representations, which are aligned to BOLD signals and evaluated with a shared voxelwise encoding model. **(b)** Spiking Transformer architecture with leaky integrate-and-fire (LIF) neurons, where continuous and spike-based features arise from the same dynamics across T time steps. **(c)** Temporal controls. Varying time steps (T) and leak base (τ) modulates model-side temporal integration and representation formation without changing model weights.

Spiking neural networks: from efficiency tools to cognitive probes. ANN-to-SNN conversion was developed for energy-efficient inference (Rueckauer et al., 2017), and spiking language models have also been trained directly, for example by distillation from BERT (Bal & Sengupta, 2024). This literature evaluates conversion by inference cost and task accuracy. We use the MA-BERT + SpikedAttention pair (Ming et al., 2023; Hwang et al., 2024), and LAS (Chen et al., 2026) for replication, as a controlled in-silico probe in which any contrast between continuous and discrete readouts is a property of the readout. Both conversions are used unmodified; our contribution is the read-only probe and the controlled variance-partitioning analysis.

3 METHOD

3.1 SAME-WEIGHT ANN/SNN SUBSTRATE

The main substrate is matrix-attention BERT (MA-BERT) paired with its SpikedAttention-converted SNN (Ming et al., 2023; Hwang et al., 2024). A standard Transformer is not directly compatible with the SpikedAttention conversion used for our matched pair, because GeLU and LayerNorm lack direct spike counterparts in that conversion family; MA-BERT therefore serves as the conversion-compatible parent. So the controlled contrast is MA-BERT against its own SNN. We use the SST-2 checkpoint that SpikedAttention distributes, unmodified. The conversion is training-free: placeholder modules in the trained ANN are calibrated against per-neuron activation maxima and replaced by spiking modules parameterised by T and τ. Linear weights are never retrained and the SNN never sees fMRI data, so any readout difference is a property of the dynamics.

Under the one-spike event code (Hwang & Kung, 2024; Hwang et al., 2024), each neuron emits at most one signed spike during a T-step forward pass. Letting $s_t \in \{-1, 0, +1\}$ be the spike at step t, the SpikedAttention continuous trace readout is

$$\mathrm{trace}_v = \sum_{t=0}^{T-1} s_t \tau^{-t}, \tag{1}$$

so earlier spikes carry larger weight, and a larger τ (a faster leak) discounts later spikes more strongly. At finite T, trace_v approximates the underlying ANN activation.

A bounded plateau could in principle be a property of one conversion family or of the matrix-attention parent, and not of spike conversion as such. To test this, we repeat the temporal analysis on a standard BERT-base-cased converted with LAS (Chen et al., 2026), a training-free conversion built on outlier-aware thresholds and hierarchically gated few-spike neurons, mechanisms absent from SpikedAttention. Checkpoint and conversion change together, so this is an independent replication that does not separate backbone from conversion.

3.2 Non-invasive readout probe

A read-only probe wraps each SNN module and caches unsigned spike counts, signed spike counts, and first-spike step per timestep, leaving the forward output numerically unchanged. From the same forward pass (Fig. 1a) we extract three candidates: `trace_v`, the leak-weighted spike sum and our continuous readout; `spike_rate`, the firing count n normalised by T and our sparse recruitment axis; and their concatenation `trace_v` $\|$ `spike_rate`, the hybrid readout that carries the main claim. Under the one-spike code $n \in \{0, 1\}$, so $r = n/T$ takes the two values $\{0, 1/T\}$.

The two readouts are views of one event sequence. A unit that fires at step t^* contributes $\pm\tau^{-t^*}$ to the trace and $1/T$ to the spike rate, and a silent unit contributes zero to both, so `spike_rate` $= \frac{1}{T}\,\mathbf{1}[$`trace_v` $\neq 0]$. The sparse readout is the network's own recruitment pattern in binary form. It records no event that the trace lacks, yet it presents those events in a format that a linear encoder cannot rebuild from the trace, because no weighted sum of `trace_v` yields its support indicator. The hybrid readout thus tests representational format with the computation held fixed, and the control of §3.3 asks whether the network's arrangement of recruited units matters. We also report `trace_v` $\|$ `signed_spike_rate` as a built-in negative control, since `trace_v` already encodes signed firing (Eq. 1). Two further readouts (`latency`, `no_spike_mask`) are catalogued in Appendix A.2.

3.3 Variance partitioning under a random-discrete control

Because a matched random-discrete block can reproduce the brain-score lift of a real sparse readout (§5.2; Feghhi et al., 2024), we test the hybrid on a per-subject paired comparison of held-out variance. Let $c =$ `trace_v`, $d =$ `spike_rate`, and $cd = c \,\|\, d$, and write R^2_m for the held-out coefficient of determination of model $m \in \{c, d, cd\}$ on its leave-one-run-out predictions. The unique discrete contribution is

$$U_d = R^2_{cd} - R^2_c, \tag{2}$$

the change in held-out variance when the discrete block is appended to the continuous trace. The random-discrete control d_{ctrl} permutes all entries of the sparse block within each run under a fixed seed and reshapes them to the original tensor. This keeps the block's value distribution, sparsity, and dimensionality, and removes the assignment of values to words and units. We then form $cd_{\text{ctrl}} = c \,\|\, d_{\text{ctrl}}$ and the per-subject candidate–control contrast

$$\Delta_i = U_d^{(i,\,\text{real})} - U_d^{(i,\,\text{ctrl})} = R^{2\,(i)}_{cd} - R^{2\,(i)}_{cd_{\text{ctrl}}}. \tag{3}$$

Both hybrids share the continuous block, the dimensionality, and the ridge pipeline, so capacity and the regularisation path cancel in this contrast, which isolates the arrangement of the recruited units. The R^2 reconstruction, the commonality decomposition, and the Pearson^2 inversion that motivates this metric are in Appendix A.

4 Experimental Setup

4.1 Naturalistic fMRI data and ROIs

We use the English subset of the Le Petit Prince multilingual naturalistic fMRI corpus (Li et al., 2022): forty-nine native English speakers listened to the same audiobook across nine runs (TR $= 2$ s; 2,856 TRs). Data are analysed in MNIColin27 space with a gray-matter mask of 181,494 voxels. Two region sets carry the claims: LanA, a probabilistic language-network mask thresholded at 0.15 (32,582 voxels) (Lipkin et al., 2022), and twelve Fedorenko parcels (LH/RH IFGorb, IFG, MFG, anterior, middle, and posterior temporal) (Fedorenko et al., 2010; Mahowald & Fedorenko, 2016), which together define a 13-ROI inferential family.

4.2 ENCODING MODEL AND BRAIN SCORE

The pipeline (Fig. 1a) feeds each sentence with prior narrative context, averages subword features to words, bins words to TRs, and applies a four-tap delay basis ($2/4/6/8$ s) to yield $X \in \mathbb{R}^{2856\times d}$ ($d = 768$ standalone, $d = 1{,}536$ hybrid). Kernel ridge regression with penalty chosen by inner two-fold cross-validation (CV) maps X to BOLD (Appendix A.3).

Outer evaluation uses leave-one-run-out cross-validation and records per-voxel held-out Pearson correlation and held-out R^2. The descriptive brain score is the across-run mean of corr^2 divided by a runwise leave-one-subject-out reliability ceiling, averaged over ROI voxels (Appendix A.4). Figures 2, 4a and 5 report brain score; the controlled analyses of Fig. 4b and Fig. 6 use held-out R^2. Subject is the only inferential unit ($n = 49$).

4.3 SPIKE-GRID SWEEPS AND INFERENCE

Five sweeps drive the Results. (i) $T \in [1, 20]$ and $\tau \in \{1.10, 1.20, 1.30, 1.40, 1.50, 1.60, 1.80\}$ on `trace_v` (140 cells; §5.1). (ii) The hybrid `trace_v` $\|$ `spike_rate` and its random-discrete control at the pre-fixed anchor cell $(T, \tau, L) = (10, 1.60, 8)$ (§5.2). The anchor lies inside the plateau, where the continuous readout is already predictive, and was fixed at the LanA group level before the control analysis, so it could not be tuned to individual listeners. (iii) $T \in \{2, 4, 6, 8, 10, 13, 16\} \times \tau \in \{1.30, 1.60\}$ on the same hybrid at layer 8 (§5.3). (iv) The matched control of (ii) at all 14 cells of (iii), at layer 8, with no selection of cells, layers, or listeners. (v) The temporal analysis on the LAS substrate ($T \in \{4, 8, 11, 13, 16\}$; layers 4, 8, 11; LanA), with endpoint, metric, layer, ROI, and listeners fixed in advance, and at $T = 13$ the matched control of (ii) with five permutation seeds and five circular-shift seeds. The pre-fixed endpoint of (v) is signed Pearson r; Appendix A.9 explains why brain score and the variance metrics are not used at that endpoint. Sweeps (iv) and (v) were run with a later version of the pipeline that z-scores BOLD within each run and builds the delay basis within runs; their absolute correlations are on a different scale, and we never compare values across the two sets. `bert-base-uncased` (Devlin et al., 2019) serves as a continuous ANN reference outside the matched pair.

The per-subject contrast Δ_i (Eq. 3) is tested by sign-flip permutation (4,096 permutations) with Benjamini–Hochberg false-discovery-rate (FDR) correction at $q < 0.05$ across the 13-ROI family (Benjamini & Hochberg, 1995). For sweeps (iv) and (v) we report 95 % bootstrap confidence intervals (CIs) over listeners (20,000 resamples), one-sided sign-flip tests (65,536 permutations), FDR correction across the 14 cells, and the number of listeners who show each effect.

5 RESULTS

§5.1 checks that the probe works and locates the $T \times \tau$ regime in which the model is aligned with cortex; §5.2 asks, inside that regime, which readout format predicts best; §5.3 asks how the hybrid advantage moves along the integration axis.

Across all 13 ROIs (Fig. 2a), the same-weight pair (MA-BERT ANN and Spiked-MA-BERT SNN at $T = 16$, $\tau = 1.40$) lands within about 3 % of the BERT-base-uncased reference, and the between-listener 95 % CI is much larger than the differences between models. At this cell, training-free conversion does not lower alignment, so the converted network can be used as a probe.

5.1 TEMPORAL INTEGRATION GOVERNS NEURAL ENCODING PERFORMANCE

Neural encoding under temporal integration control. Sweeping the full $T \in [1, 20] \times \tau \in \{1.10, \ldots, 1.80\}$ grid (140 cells; Fig. 2b, best layer per cell) shows how the unfolding of the converted network shapes encoding from `trace_v`. At $T \leq 3$, brain score is markedly lower at every τ (0.34 to 0.47). It then rises and levels off: for $\tau \geq 1.4$ it reaches 0.49 to 0.52 at $T \approx 9$–11, the level of the parent ANN in Fig. 2a, and stays there up to $T = 20$, with the highest scores at large τ. The converted SNN therefore needs about ten steps to recover the brain alignment of its parent model, and further steps bring no gain. Table 9 shows where these boundaries come from: at $T \leq 4$ almost no layer-8 unit fires, so the trace (Eq. 1) is close to zero; the fraction of firing units passes one half between $T = 8$ and $T = 10$ under $\tau = 1.30$ and between $T = 4$ and $T = 6$ under $\tau = 1.60$,

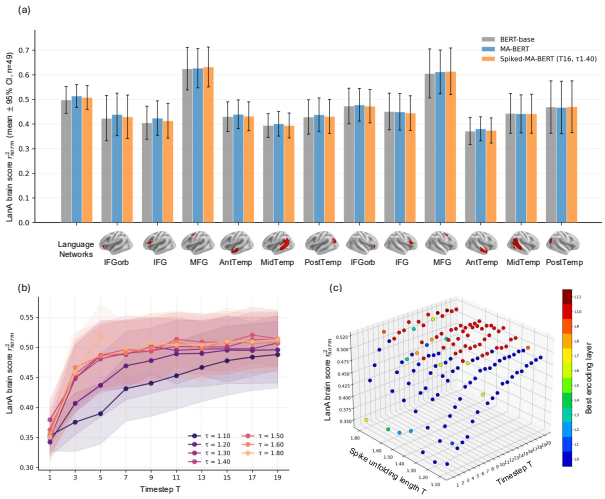

Figure 2: **Backbone preservation and bounded temporal integration.** **(a)** Same-weight MA-BERT / Spiked-MA-BERT pair preserves language-network alignment (13 ROIs, mean $\pm$ 95% CI, $n = 49$). **(b)** Brain score of the `trace_v` readout across $T \times \tau$ on LanA, reporting the best layer per cell; note that the vertical axis starts at 0.30. **(c)** Joint $T \times \tau$ landscape (brain score as height), with colour indicating the best encoding layer.

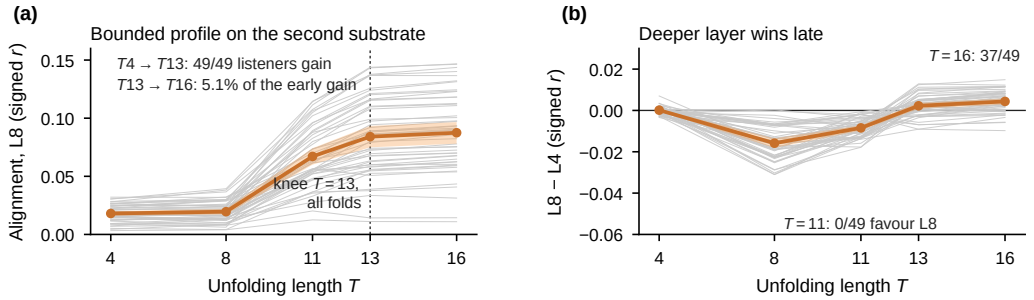

Figure 3: **The temporal profile replicates on a second substrate.** BERT-base-cased converted with LAS, LanA, $n = 49$. Thin lines and dots are listeners; thick lines are group means with 95% bootstrap CIs. **(a)** Alignment at layer 8 rises with T and stabilises by $T = 13$. **(b)** Layer 8 minus layer 4: the deeper layer overtakes the shallower one only late in the unfolding.

and the plateau begins where most units have fired. Because the text input is identical in every cell, this bounded plateau is a property of the model's encoding performance; T and τ index model-side integration and are not a measurement of a cortical window.

Layer preference shifts with integration. Figure 2c colours the $T \times \tau$ landscape by best encoding layer: shallow layers (L_0–L_2) dominate at $T < 10$, and the preferred layer moves to L_9–L_{10} as the system approaches the plateau.

Replication on a second substrate. The same profile appears on BERT-base-cased converted with LAS (Fig. 3; layer 8, LanA, signed Pearson r, 95 % bootstrap CIs over 49 listeners). Alignment gains $+0.0661$ $[+0.0579, +0.0743]$ from $T = 4$ to $T = 13$ in $49/49$ listeners and only $+0.0034$ $[+0.0028, +0.0039]$ from $T = 13$ to $T = 16$, 5.1 % of the early gain. A bounded piecewise form predicts held-out T better than an unbounded log-linear one (leave-one-T-out SSE 0.0032 against 0.0053; knee at $T = 13$ in every fold; preferred in $40/49$ listeners). The layer shift also reproduces (Fig. 3b): the advantage of layer 8 over layer 4 rises by $+0.0128$ $[+0.0105, +0.0150]$ between $T = 11$ and $T = 16$ ($43/49$), from $0/49$ listeners favouring layer 8 at $T = 11$ to $37/49$ at $T = 16$. Over the stable range $T = 8$ to 16, the early contrast is positive in $49/49$ listeners under four metrics

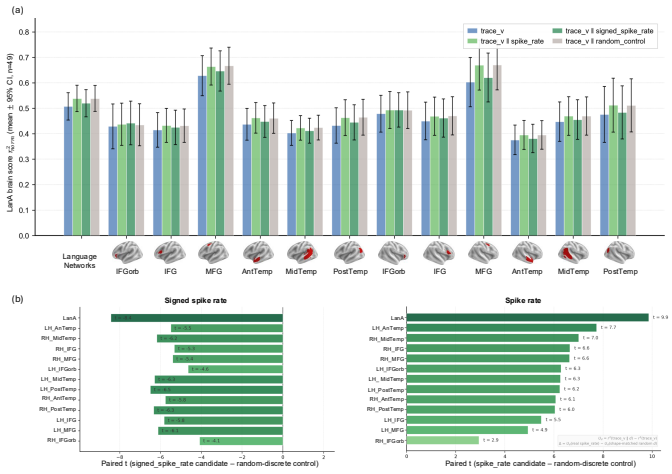

Figure 4: **Hybrid coding under a random-discrete control.** **(a)** Brain score across ROIs for continuous, hybrid, and random-discrete representations. **(b)** Paired t of the candidate–control contrast Δ_i on held-out R^2 across ROIs. Left: `signed_spike_rate` vs matched random control. Right: `spike_rate` vs matched random control.

(Appendix A.9). The plateau and the layer shift therefore do not depend on SpikedAttention or on the matrix-attention parent; both parents are encoder-only BERT models, and checkpoint and conversion change together. The plateau fixes the regime for what follows.

5.2 Hybrid coding under a random-discrete control

Inside the plateau, we ask which readout format predicts held-out responses best, comparing at the pre-fixed anchor cell ($T = 10$, $\tau = 1.60$, $L = 8$): `trace_v`, the sparse axis `spike_rate` and its signed variant `signed_spike_rate`, and their concatenations with `trace_v`, all read from the same forward pass of the post-feed-forward output LIF module.

Brain score cannot separate real from random. Across LanA and the twelve Fed parcels (Fig. 4a), both real hybrids exceed `trace_v` alone on every ROI by $+0.027$ in ceiling-normalised brain score. The matched random-discrete hybrid (§3.3) reaches the same $+0.027$ lift: a per-subject paired t-test of real minus matched random gives $|t| < 1.7$ on every ROI, with $0/13$ FDR-significant. Squared correlation rewards the wider spread of fold-wise correlations that either block produces (Appendix A.7; Feghhi et al., 2024), so brain score cannot tell a real hybrid from a matched random one.

Held-out variance separates them. We therefore compare the two hybrids on held-out R^2, paired across subjects and tested by sign-flip permutation (Eq. 3; Fig. 4b). The real `spike_rate` block predicts held-out responses better than the matched random block on every one of the 13 ROIs, with paired $t = +2.94$ (RH IFGorb) to $+9.86$ (LanA), $13/13$ FDR-significant at $q < 0.05$. The `signed_spike_rate` block goes the other way: paired $t = -4.1$ (RH IFGorb) to -8.4 (LanA), $13/13$ FDR-significant in the negative direction, so the matched random block outperforms the real signed candidate on every ROI. Repeating the control at all fourteen cells of the §5.3 grid (sweep iv) reproduces the anchor result ($+0.000144$ $[+0.000066, +0.000227]$, $33/49$ listeners, $q = 0.0017$); at the anchor it is also FDR-positive under explained variance ($+0.000141$, $q = 0.0015$) and has the same sign under signed r without significance ($+0.000211$, $29/49$, $q = 0.065$; Table 7). The contrast is FDR-positive in six cells, all in the mid-to-long unfolding range ($T = 10$ to 16 at $\tau = 1.30$, $T = 6$ to 10 at $\tau = 1.60$), and negative or non-significant elsewhere (Appendix A.10). The same contrast holds on the LAS substrate at $T = 13$ under held-out R^2 ($+0.00026$ $[+0.00007, +0.00046]$, $30/49$ listeners, five permutation seeds), and against circularly shifted blocks that keep the real block's temporal autocorrelation ($+0.00023$, $31/49$; Appendix A.9).

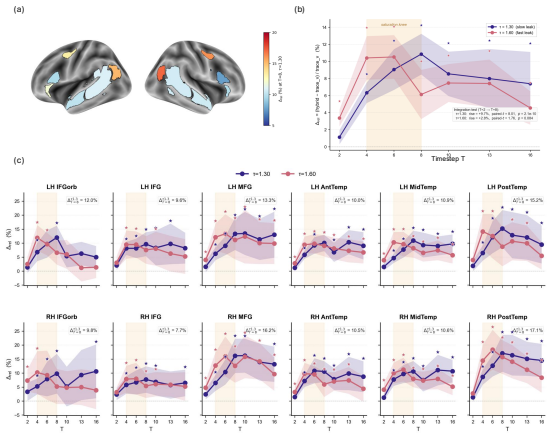

Figure 5: **Relative hybrid lift across temporal integration (descriptive, brain score). (a)** Δ_{rel} across Fed language parcels at $T = 8$, $\tau = 1.30$. **(b)** LanA trajectory of Δ_{rel} across T under two τ settings. **(c)** Per-parcel trajectories. Δ_{rel} is not controlled against the matched random block; the controlled profile is in Fig. 6.

Signed and unsigned recruitment. The divergence follows from the substrate: `trace_v` already carries the signed firing of each unit (Eq. 1), so the signed block repeats what the trace holds, whereas unsigned `spike_rate` encodes which units were recruited regardless of polarity, a function of the same computation that a linear encoder cannot reach from the trace (§3.2). The choice of metric matters: a Pearson-r^2 version of the same test reverses the conclusion ($11/13$ ROIs favour the random control; Appendix A.7), so the hybrid result is scoped to held-out R^2. Whether the advantage depends on where along the integration axis it is measured is the question of the next section.

5.3 Hybrid coding across integration and language parcels

§5.2 showed that the hybrid format has an advantage at one cell inside the plateau. Here we ask how that advantage moves along the model-side integration axis, first descriptively over the whole grid and then under the matched control. For the descriptive view we track the per-parcel *relative* hybrid lift

$$\Delta_{\text{rel}}(T, \tau) \;=\; \frac{B_{cd}(T, \tau) - B_c(T, \tau)}{B_c(T, \tau)}, \tag{4}$$

where B_m is the ROI brain score of model m (Eq. 6), computed per listener and then averaged. Brain score does not separate a real block from a matched random one (§5.2), so we read Δ_{rel} as a description and carry the inferential claim on the controlled contrast below.

Spatial variation and per-parcel trajectories. Figure 5a colours each Fed parcel by $\Delta_{\text{rel}}(T = 8, \tau = 1.30)$. The relative lift is largest in bilateral posterior-temporal cortex (RH PostTemp $+17.1\,\%$, LH PostTemp $+15.2\,\%$) and middle-frontal cortex (RH MFG $+16.2\,\%$, LH MFG $+13.3\,\%$) and smallest in inferior-frontal cortex (RH IFG $+7.7\,\%$, RH IFGorb $+9.8\,\%$). In every parcel the trajectory rises, peaks, and declines, with a significant $T = 2 \rightarrow T = 8$ rise at $\tau = 1.30$ and a maximum at a smaller T under the faster leak (Fig. 5c; Table 3), so the heterogeneity is in amplitude. This pattern is descriptively consistent with regions previously identified as long-timescale integrators in the temporal-receptive-window hierarchy (Lerner et al., 2011; Honey et al., 2012); we report it as a correspondence of the uncontrolled lift and do not attribute it to specific linguistic operations.

Temporal profile on LanA. On LanA, Δ_{rel} is non-monotonic (Fig. 5b). Under the slow leak $\tau = 1.30$, the lift climbs from $+1.1\,\%$ at $T = 2$ to $+10.85\,\%$ at $T = 8$ and falls back to $+7.4\,\%$ by $T = 16$; under the faster leak $\tau = 1.60$, it peaks earlier, at $T = 4$–6 near $+10.5\,\%$, and decreases to $+4.5\,\%$ at $T = 16$. The rising arm at $\tau = 1.30$ is significant (paired $t = 8.01$, $p = 2.1 \times 10^{-10}$, $46/49$ listeners), and Δ_{rel} is above zero in 13 of the 14 cells after FDR correction across the LanA series. The faster leak moves the whole arc to earlier T without raising its peak, which is consistent

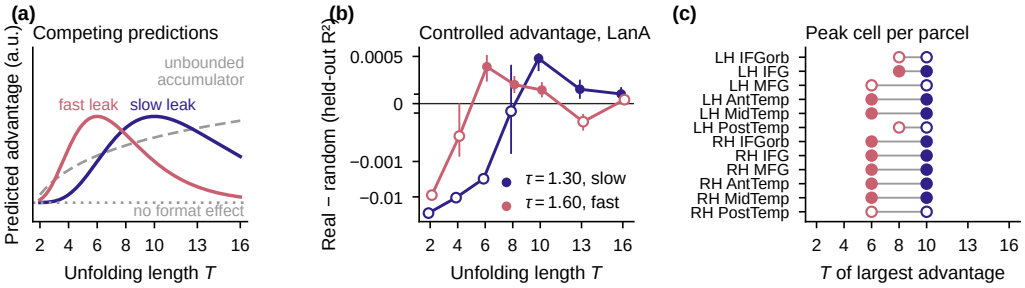

Figure 6: **Temporal profile of the hybrid advantage under the matched control. (a)** Schematic of three competing predictions (no data): an unbounded accumulator, no format effect, and a finite window whose peak moves earlier under faster leak. **(b)** Real minus matched random hybrid on LanA (held-out R^2, mean and 95% bootstrap CI, $n = 49$, sweep iv); the axis is linear within $\pm 5 \times 10^{-4}$ and logarithmic beyond. **(c)** Cell with the largest mean advantage per Fed parcel and leak. Filled markers are FDR-positive.

with what would be expected under an integration-window mechanism. The decline at large T is a property of the ratio: the continuous baseline keeps growing slowly into the plateau while the hybrid increment stops at the knee. A constrained finite-window form fitted to these trajectories is supported component by component (early rise in $46/49$ listeners at $\tau = 1.30$, post-peak decline in $35/49$ at $\tau = 1.60$, earlier fitted peak under the faster leak in $26/49$) and has no cross-validated advantage over a monotone form as a whole (Appendix A.10).

Temporal profile under the matched control. The matched control gives the same picture with the random block subtracted (Fig. 6b; sweep iv, held-out R^2). The advantage of the real over the random hybrid rises, peaks, and declines in T, with the peak at $T = 10$ under $\tau = 1.30$ ($+0.00048$, $42/49$ listeners) and at $T = 6$ under $\tau = 1.60$ ($+0.00039$, $39/49$), both $q = 1.1 \times 10^{-4}$. It is negative at short unfolding, before the continuous readout reaches the plateau (§5.1). The two peak cells are the first cells at which more than half of the layer-8 units fire (Table 9). Nearly all units are silent below that point and nearly all fire above it, so the recruitment pattern varies most near the crossing, which the faster leak reaches sooner. The finite window therefore has a model-side reading: the sparse readout carries the most variance where recruitment is neither sparse nor saturated. In the twelve Fed parcels the largest advantage falls at $T = 10$ under the slow leak in every parcel and at $T = 6$ or 8 under the fast leak, so the peak is earlier under faster leak in $12/12$ parcels (Fig. 6c); eight parcels have at least one FDR-positive cell at each leak (Appendix A.10). Of the three predictions in Fig. 6a, the controlled profile is closest to the finite window.

6 Conclusion

We introduced a same-weight ANN↔SNN substrate as a controlled probe of representational format and temporal integration. On naturalistic fMRI from 49 listeners of one audiobook, brain alignment saturates inside a bounded $T \times \tau$ plateau at the level of the parent model, where the preferred layer shifts from shallow to deep, and the profile replicates in every listener on a second backbone and conversion family. Inside the plateau, the hybrid continuous–sparse readout predicts held-out responses better than a matched random-discrete control on 13/13 language-network ROIs, while a redundant signed variant does not, and this advantage peaks earlier under faster leak in all twelve language parcels. The findings are consistent with a hybrid continuous–sparse code that is useful only inside a finite integration window, and they suggest that finite temporal integration may be among the inductive biases that bring learned language representations closer to those of the human language network. T and τ are model-side variables, so the plateau and the window describe the converted model, not a cortical timescale, and every result comes from one narrative (Appendix B).

AI USE STATEMENT

In this work, we used generative AI tools for one task with required disclosure under the ICLR 2027 AI Policy for Authors: implementing part of the analysis code. We have not used generative AI tools to generate synthetic data, to develop the theoretical models or the conceptual framework, to propose or refine hypotheses, to design or provide feedback on the research methodology or experiments, to clean or reformat the dataset, or to interpret the results; formulating or proving mathematical claims, translation, and qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for tasks with recommended disclosure: editing the manuscript to improve readability; debugging and reorganising existing analysis code; writing and submitting batch job scripts on the computing cluster and searching for information and identifying relevant literature. We have reviewed all AI-assisted work. All AI-implemented, AI-debugged or AI-reorganised code and job scripts were manually checked, run and tested by the authors; AI-edited text was checked by the authors against the reported results and the source files behind them; the reported numbers were checked against the saved analysis outputs; every figure was checked against its source data. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work analyses an existing, anonymised, public fMRI corpus, the English subset of *Le Petit Prince* (OpenNeuro ds003643, CC0; Li et al., 2022), together with publicly released pretrained language-model weights. We collected no new human-subjects data; participant recruitment, consent and institutional review are documented by the dataset authors. The main risk we see is over-interpretation: spike-derived model variables can sound biologically more direct than slow fMRI evidence warrants. We therefore treat spatial correspondences as descriptive and restrict inferential claims to subject-level effects.

REPRODUCIBILITY STATEMENT

The substrate, readouts and variance-partitioning protocol are specified in §3.1–§3.3; subjects, runs, ROIs, the delay basis, the ridge configuration, the sweep grids and the inference policy are given in §4.1–§4.3, with full formulas and the ridge penalty grid in Appendix A. The fMRI corpus (Li et al., 2022), MA-BERT (Ming et al., 2023) and SpikedAttention (Hwang et al., 2024) are publicly available, and we use the published weights and conversion procedure unchanged. The encoding pipeline, the seeded random-discrete control builder, the held-out R^2 and commonality routines, the paired sign-flip test, per-figure reproduction scripts and SLURM templates are provided in the supplementary material.

REFERENCES

Malyaban Bal and Abhronil Sengupta. Spikingbert: Distilling bert to train spiking language models using implicit differentiation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 10998–11006, 2024.

Christopher Baldassano, Janice Chen, Asieh Zadbood, Jonathan W Pillow, Uri Hasson, and Kenneth A Norman. Discovering event structure in continuous narrative perception and memory. *Neuron*, 95 (3):709–721, 2017.

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

Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications biology*, 5(1):134, 2022.

Long Chen, Xiaotian Song, and Yanan Sun. Las: Loss-less ann-snn conversion for fully spike-driven large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 1730–1738, 2026.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 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. 4171–4186, 2019.

Evelina Fedorenko, Po-Jang Hsieh, Alfonso Nieto-Castañón, Susan Whitfield-Gabrieli, and Nancy Kanwisher. New method for fmri investigations of language: defining rois functionally in individual subjects. *Journal of neurophysiology*, 104(2):1177–1194, 2010.

Ebrahim Feghhi, Nima Hadidi, Bryan Song, Idan A Blank, and Jonathan C Kao. What are large language models mapping to in the brain? a case against over-reliance on brain scores. *arXiv preprint arXiv:2406.01538*, 2024.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. Shared computational principles for language processing in humans and deep language models. *Nature neuroscience*, 25(3):369–380, 2022.

Uri Hasson, Eunice Yang, Ignacio Vallines, David J Heeger, and Nava Rubin. A hierarchy of temporal receptive windows in human cortex. *The Journal of neuroscience*, 28(10):2539–2550, 2008.

Christopher J Honey, Thomas Thesen, Tobias H Donner, Lauren J Silbert, Chad E Carlson, Orrin Devinsky, Werner K Doyle, Nava Rubin, David J Heeger, and Uri Hasson. Slow cortical dynamics and the accumulation of information over long timescales. *Neuron*, 76(2):423–434, 2012.

Alexander G Huth, Wendy A De Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532 (7600):453–458, 2016.

Sangwoo Hwang and Jaeha Kung. One-spike snn: Single-spike phase coding with base manipulation for ann-to-snn conversion loss minimization. *IEEE Transactions on Emerging Topics in Computing*, 13(1):162–172, 2024.

Sangwoo Hwang, Seunghyun Lee, Dahoon Park, Donghun Lee, and Jaeha Kung. Spikedattention: Training-free and fully spike-driven transformer-to-snn conversion with winner-oriented spike shift for softmax operation. *Advances in Neural Information Processing Systems*, 37:67422–67445, 2024.

Yulia Lerner, Christopher J Honey, Lauren J Silbert, and Uri Hasson. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *The Journal of neuroscience*, 31 (8):2906–2915, 2011.

Jixing Li, Shohini Bhattasali, Shulin Zhang, Berta Franzluebbers, Wen-Ming Luh, R Nathan Spreng, Jonathan R Brennan, Yiming Yang, Christophe Pallier, and John Hale. Le petit prince multilingual naturalistic fmri corpus. *Scientific data*, 9(1):530, 2022.

Benjamin Lipkin, Greta Tuckute, Josef Affourtit, Hannah Small, Zachary Mineroff, Hope Kean, Olessia Jouravlev, Lara Rakocevic, Brianna Pritchett, Matthew Siegelman, et al. Probabilistic atlas for the language network based on precision fMRI data from> 800 individuals. *Scientific data*, 9 (1):529, 2022.

Kyle Mahowald and Evelina Fedorenko. Reliable individual-level neural markers of high-level language processing: A necessary precursor for relating neural variability to behavioral and genetic variability. *NeuroImage*, 139:74–93, 2016.

Neo Wei Ming, Zhehui Wang, Cheng Liu, Rick Siow Mong Goh, and Tao Luo. Ma-bert: Towards matrix arithmetic-only bert inference by eliminating complex non-linear functions. In *ICLR*, 2023.

Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fmri. *Neuroimage*, 56(2):400–410, 2011.

Alexandre Pasquiou, Yair Lakretz, John T Hale, Bertrand Thirion, and Christophe Pallier. Neural language models are not born equal to fit brain data, but training helps. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 17499–17516. PMLR, 17–23 Jul 2022.

Bodo Rueckauer, Iulia-Alexandra Lungu, Yuhuang Hu, Michael Pfeiffer, and Shih-Chii Liu. Conversion of continuous-valued deep networks to efficient event-driven networks for image classification. *Frontiers in neuroscience*, 11:682, 2017.

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.

A Method and experimental-setup details

A.1 Notation

Table 1: Symbols used in the main text.

Symbol	Meaning
T	unfolding length of the converted network (timesteps per pass); model-side
τ	leak base of the leaky integrate-and-fire units (Eq. 1); larger τ is a faster leak
L	encoder layer from which readouts are taken
c, d, cd	continuous readout `trace_v`, sparse readout `spike_rate`, and their concatenation
d_{ctrl}, cd_{ctrl}	matched random-discrete block and the control hybrid (§3.3)
R^2_m	held-out R^2 of model m on leave-one-run-out predictions (Eq. 7)
U_d	unique discrete contribution $R^2_{cd} - R^2_c$ (Eq. 2)
Δ_i	per-listener candidate–control contrast $R^{2\,(i)}_{cd} - R^{2\,(i)}_{cd_{\text{ctrl}}}$ (Eq. 3)
B_m	ceiling-normalised brain score of model m (Eq. 6)
Δ_{rel}	relative hybrid lift $(B_{cd} - B_c)/B_c$ (Eq. 4); descriptive

A.2 Edge readouts

Two further readouts are catalogued for completeness and are not part of the candidate family that drives the main claims:

- `latency`: first-spike step normalised to $[0, 1]$, which captures spike timing within a T-step pass;
- `no_spike_mask`: silent-neuron indicator (1 if the neuron does not fire across the T steps, else 0).

Under the one-spike code, `no_spike_mask` $= 1 - T \cdot$ `spike_rate`, so the two span the same column space once the design matrix is standardised, and we treat `spike_rate` as the single recruitment readout. Both edge readouts score below `trace_v` when used standalone, and we exclude them from the hybrid test and the integration sweeps.

A.3 Feature extraction details

For each sentence, the model receives the current sentence plus available prior context from the same narrative section, padded or truncated to 128 tokens; hidden states are sliced back to the current sentence and concatenated across the nine runs. The text input is therefore identical in every (T, τ) cell. Subword features are averaged to word level by word index, each word is assigned to a TR by $\lfloor \text{onset}/\text{TR} \rfloor$, and within-TR words are mean-pooled. The nine runs stack to a $2{,}856 \times d$ matrix per readout, with per-run TR counts $(282, 298, 340, 303, 265, 343, 325, 292, 368)$ for runs 1–9 respectively. We model the hemodynamic lag with a four-tap FIR-like delay basis using $[1, 2, 3, 4]$ TR shifts ($2/4/6/8$ s at TR $= 2$ s) and standardise the delayed matrix before kernel ridge regression. The α grid and kernel-ridge rationale are in §A.6.

A.4 BRAIN-SCORE AND NOISE-CEILING: FULL FORMULAS

For subject i, run r, voxel v, the runwise leave-one-subject-out reliability ceiling is

$$\text{ceiling}_{i,v} = \frac{1}{9}\sum_{r=1}^{9} \text{corr}\left(y_{i,r,v},\ \frac{1}{48}\sum_{j\neq i} y_{j,r,v}\right)^2, \tag{5}$$

and the voxel-level brain score is

$$\text{brain_score}_{i,v} = \frac{\frac{1}{9}\sum_{r=1}^{9} \text{corr}(y_{i,r,v}, \hat{y}_{i,r,v})^2}{\text{ceiling}_{i,v}}. \tag{6}$$

ROI scores are voxel-mean ratios; subject is the only inferential unit ($n = 49$).

A.5 HELD-OUT R^2 RECONSTRUCTION AND COMMONALITY DECOMPOSITION

We reconstruct the classic held-out R^2 from saved leave-one-run-out predictions without refitting:

$$R^2_m = 1 - \frac{\sum_t (y_t - \hat{y}_{m,t})^2}{\sum_t (y_t - \bar{y})^2}, \quad m \in \{c, d, cd\}. \tag{7}$$

The commonality decomposition for the candidate hybrid is

$$U_c = R^2_{cd} - R^2_d, \quad U_d = R^2_{cd} - R^2_c, \quad S = R^2_c + R^2_d - R^2_{cd}, \tag{8}$$

with the algebraic guard $U_c + U_d + S = R^2_{cd}$. The same decomposition is computed in parallel on $(c, d_{\text{ctrl}}, cd_{\text{ctrl}})$ to obtain the control-side $U_d^{(\text{ctrl})}$ used in Eq. 3. The control permutation is drawn once per run and layer from a fixed seed; after permutation the sorted values of the block are identical to those of the real block, which is the matching we rely on.

A.6 RIDGE REGRESSION CONFIGURATION

The implementation uses *kernel* ridge cross-validation in place of primal ridge because the delayed feature dimension exceeds the number of time points in the hybrid setting ($1{,}536 \times 4 > 2{,}856$). The penalty α is selected by inner two-fold cross-validation over the wide grid

$$\alpha \in \{10^0, 10^1, 10^2, \ldots, 10^{19}\}.$$

The penalty is selected separately for every voxel. The two inner folds are the first and second halves of the training TRs in time order for sweeps (i)–(iii), and two groups of whole runs for sweeps (iv) and (v). The grid spans 20 decades because the raw readouts differ sharply in magnitude (continuous trace_v is much larger than $\text{spike_rate} \in \{0, 1/T\}$), so the optimum can sit several decades apart between cells. In every cell we report, the selected α is interior to the grid; no cell saturates the 10^0 or 10^{19} endpoint.

A.7 WHERE THE PEARSON^2 INVERSION COMES FROM

We selected classic held-out R^2 together with U_d as the metric for the hybrid claim only after observing an empirical inversion under a Pearson^2 proxy. At the pre-fixed anchor cell $(T, \tau, L) = (10, 1.60, 8)$, running the same paired sign-flip test on U_d with Pearson^2 in place of classic held-out R^2 inverts the conclusion: the random-discrete control numerically beats the real candidate on $11/13$ ROIs ($t = -1.35$ to -8.53, subject as the inferential unit), so the inversion is systematic across ROIs.

Table 2 shows where the inversion comes from. On LanA at the anchor, appending either 768-dimensional block lowers the mean fold-wise correlation relative to `trace_v` alone and raises the mean squared correlation. Added directions widen the dispersion of fold-wise correlations, and squaring turns that dispersion into positive mass whatever the block contains, so the real and the matched random hybrid are indistinguishable at the third decimal of brain score while both exceed the continuous baseline. Classic held-out R^2 is sensitive to the scale and bias of the prediction and separates the two hybrids on the same data (§5.2). This diagnostic was run at one cell and on one ROI family; per-listener inversion counts and sensitivity to preprocessing were not tested.

Table 2: Correlation, squared correlation, and brain score on LanA at the anchor cell ($n = 49$, leave-one-run-out folds). Brain score is ceiling-normalised squared correlation (Eq. 6).

Representation	mean fold r	mean fold r^2	brain score
`trace_v`	0.008620	0.009916	0.5071
`trace_v`‖`spike_rate` (real)	0.008122	0.010374	0.5384
`trace_v`‖random-discrete (control)	0.008496	0.010377	0.5385

A.8 Per-parcel rise / peak / saturation summary

Table 3 reports, for each Fed parcel and $\tau \in \{1.30, 1.60\}$, the $T = 2 \to T = 8$ rise in Δ_{rel}, its paired-t statistic across $n = 49$ subjects with FDR-BH significance, the relative lift at $T = 8$, $\Delta_{\mathrm{rel}}(T = 8, \tau)$, and the timestep $T_{\mathrm{sat}} \in \{2, 4, 6, 8, 10, 13, 16\}$ at which the group-mean Δ_{rel} is largest. T_{sat} is the argmax over seven grid points and carries no uncertainty estimate.

Table 3: Per-parcel summary of the relative hybrid lift Δ_{rel} on the Fed parcels (Fig. 5c; brain score, descriptive). $*$ = FDR-BH significant at $q < 0.05$ across $n = 49$ subjects (sign-flip permutation, 4096 perms).

Parcel	τ	rise (%)	t	$\Delta_{\mathrm{rel}}^{T=8}$ (%)	T_{sat}	FDR
LH AntTemp	1.3	8.84	5.65	10.04	8	*
LH AntTemp	1.6	6.25	3.46	9.04	4	*
LH IFG	1.3	7.58	3.28	9.62	8	*
LH IFG	1.6	4.57	1.76	7.56	4	
LH IFGorb	1.3	10.64	5.17	12.00	8	*
LH IFGorb	1.6	4.08	1.73	6.59	4	
LH MFG	1.3	11.77	3.61	13.30	8	*
LH MFG	1.6	7.11	2.07	11.12	4	
LH MidTemp	1.3	9.38	4.85	10.89	8	*
LH MidTemp	1.6	4.27	2.33	8.17	4	
LH PostTemp	1.3	13.32	3.77	15.17	8	*
LH PostTemp	1.6	4.79	1.28	8.74	4	
RH AntTemp	1.3	9.08	4.72	10.55	6	*
RH AntTemp	1.6	2.73	1.07	5.96	4	
RH IFG	1.3	5.33	2.07	7.68	8	*
RH IFG	1.6	2.26	1.01	5.41	4	
RH IFGorb	1.3	6.47	2.27	9.79	8	*
RH IFGorb	1.6	-2.16	-0.66	5.19	4	
RH MFG	1.3	13.77	4.35	16.17	8	*
RH MFG	1.6	7.81	2.36	12.59	6	
RH MidTemp	1.3	9.36	4.93	10.61	8	*
RH MidTemp	1.6	3.84	2.52	7.91	6	
RH PostTemp	1.3	15.73	4.80	17.07	8	*
RH PostTemp	1.6	12.96	4.41	15.71	6	*

A.9 Replication on the second substrate: metrics and numerical range

Table 4 repeats the early contrast of the LAS replication (sweep v) under four summaries. The primary endpoint of the replication, fixed in advance, is signed Pearson r over $T = 4$ to $T = 13$ (§5.1). Brain score is not reported on this substrate because the noise ceiling of Eq. 5 was computed under the original pipeline and was not recomputed under the later one; the three variance-based summaries were added afterwards as sensitivity checks and are reported over $T = 8$ to $T = 13$. The shorter range is needed because the LAS substrate has a prediction-scale failure at $T = 4$: the correlation stays positive (mean signed $r = 0.0180$) while the pooled held-out R^2 is -5.47×10^5, and the subject-wise $\log_{10}$ ratio of pooled squared error at $T = 4$ over $T = 8$ is 5.70 (95 % CI $[5.66, 5.75]$, $49/49$ listeners). Correlation does not register this scale error, so the variance-based agreement in Table 4 is a descriptive sensitivity result over the stable range and does not replace the pre-fixed endpoint.

Table 4: Early contrast $T13 - T8$ on the LAS substrate under four summaries (layer 8, LanA, $n = 49$; signed r in correlation units, the others as variance fractions; 95% bootstrap CIs over listeners). The last column is the ratio of the late change ($T16 - T13$) to the early change.

Summary	$T13 - T8$	95% CI	Listeners	Late/early
Signed Pearson r	$+0.0645$	$[+0.0566, +0.0724]$	49/49	0.052
Voxel-mean held-out R^2	$+0.0188$	$[+0.0165, +0.0212]$	49/49	0.045
Pooled held-out R^2	$+0.0212$	$[+0.0187, +0.0239]$	49/49	0.044
Pooled explained variance	$+0.0212$	$[+0.0186, +0.0239]$	49/49	0.044

Hybrid contrast on the second substrate. Table 5 repeats the candidate–control contrast of §5.2 on the LAS substrate at $T = 13$, layer 8, LanA, with two null families of five seeds each: the permuted block of §3.3, and a circularly shifted block that keeps the temporal autocorrelation of the real block and moves only its alignment to the words. Under both variance metrics the real hybrid beats the mean null of either family; under signed r it does not, as at the MA-BERT anchor (Table 7). The listener counts (30 to 31 of 49) are below those of the MA-BERT anchor.

Table 5: Real hybrid minus the mean of five null hybrids on the LAS substrate ($T = 13$, layer 8, LanA, $n = 49$; five seeds per null family; 95% bootstrap CIs over listeners).

Null family	Metric	Mean	95% CI	Listeners
Permuted block	Pooled held-out R^2	$+0.000255$	$[+0.000073, +0.000458]$	30/49
Permuted block	Pooled explained variance	$+0.000252$	$[+0.000072, +0.000451]$	30/49
Permuted block	Signed Pearson r	-0.000324	$[-0.000550, -0.000104]$	22/49
Circularly shifted block	Pooled held-out R^2	$+0.000230$	$[+0.000057, +0.000419]$	31/49
Circularly shifted block	Pooled explained variance	$+0.000228$	$[+0.000053, +0.000418]$	30/49
Circularly shifted block	Signed Pearson r	-0.000433	$[-0.000668, -0.000198]$	16/49

A.10 Matched control across the temporal grid

Complete 14-cell grid. Table 6 reports the real-minus-matched-random contrast at all fourteen $T \times \tau$ cells of sweep (iv): MA-BERT with SpikedAttention, layer 8, LanA, 49 listeners, the seed-0 permutation, no PCA, and every cell retained. Under held-out R^2 and explained variance the contrast rises, peaks, and declines in T, and it is FDR-positive in six cells. Under signed Pearson r the profile differs: the contrast is largest at short unfolding for both leaks and FDR-positive in four cells, so the earlier peak under faster leak is a property of the variance-based metrics. Table 7 gives the three metrics at the pre-fixed anchor.

Per-parcel expansion. We evaluated the same contrast under LanA and each of the twelve Fed parcels from the stored per-voxel moments of sweep (iv), with the ROI family, cells, metrics, and summaries fixed before any per-parcel value was computed; the LanA values reproduce those of Table 6 exactly. Table 8 lists, for each ROI and leak, the cell with the largest mean contrast and the FDR-positive cells. The largest advantage falls at $T = 10$ under $\tau = 1.30$ in every parcel and at $T = 6$ or 8 under $\tau = 1.60$, so the peak is strictly earlier under the faster leak in $12/12$ parcels. 8 of the 12 parcels have at least one FDR-positive cell at each leak; we had set $10/12$ as the threshold for moving a controlled per-parcel panel into the main text, so Fig. 5a and 5c remain descriptive. The parcel ranking of the controlled contrast does not follow the ranking of the relative lift in Fig. 5a (Spearman $\rho = -0.49$ across the 12 parcels at $T = 8$, $\tau = 1.30$), which is a further reason to read that map as a description of the uncontrolled lift. Under signed r the peak is not systematically earlier under the faster leak, in line with Table 6. The complete per-parcel grid under all three metrics accompanies the supplementary material.

Functional forms for the relative lift. We fitted the submitted Δ_{rel} trajectories (49 listeners, seven T, both τ, LanA, the twelve Fed parcels, and their mean; 14 ROI series) with a constrained finite-window form and a monotone log-linear form, compared out of sample with the knee selected inside nested leave-one-T-out folds. The finite-window form has a cross-validated advantage whose

Table 6: Real minus matched random hybrid at every cell of the temporal grid (LanA, layer 8, $n = 49$). Values are means over listeners in units of 10^{-4}; n^+ is the number of listeners with a positive contrast; * marks cells that are FDR-positive across the 14 cells of that metric ($q < 0.05$, one-sided sign-flip test).

		Pooled held-out R^2			Explained variance		Signed r	
τ	T	mean	95% CI	n^+	mean	n^+	mean	n^+
1.30	2	-287.46	$[-347.48, -233.56]$	0	-278.27	0	$+86.18^*$	39
1.30	4	-105.20	$[-135.00, -82.54]$	0	-101.13	0	$+99.38^*$	45
1.30	6	-30.78	$[-37.19, -24.97]$	0	-29.34	0	$+72.74^*$	44
1.30	8	-0.80	$[-6.10, +4.11]$	29	-0.76	29	-10.75	18
1.30	10	$+4.78^*$	$[+3.44, +6.16]$	42	$+4.72^*$	41	$+1.48$	28
1.30	13	$+1.53^*$	$[+0.51, +2.54]$	29	$+1.50^*$	29	$+1.44$	28
1.30	16	$+1.02^*$	$[+0.28, +1.75]$	32	$+1.00^*$	32	$+2.22$	32
1.60	2	-89.05	$[-105.62, -74.05]$	0	-85.37	0	$+14.14$	29
1.60	4	-3.44	$[-7.32, +0.03]$	22	-2.97	22	$+48.75^*$	38
1.60	6	$+3.91^*$	$[+2.34, +5.48]$	39	$+3.86^*$	39	-0.20	24
1.60	8	$+2.01^*$	$[+1.13, +2.92]$	33	$+1.96^*$	33	$+2.38$	32
1.60	10	$+1.44^*$	$[+0.66, +2.27]$	33	$+1.41^*$	33	$+2.11$	29
1.60	13	-1.92	$[-2.87, -1.10]$	12	-1.93	12	-0.62	23
1.60	16	$+0.42$	$[-0.02, +0.88]$	27	$+0.42$	27	-0.47	21

Table 7: Real minus matched random hybrid at the pre-fixed anchor cell $(T, \tau, L) = (10, 1.60, 8)$ under three metrics (LanA, $n = 49$; q is the FDR-adjusted one-sided sign-flip p across the 14 cells).

Metric	Mean	95% CI	Listeners	q
Pooled held-out R^2	$+0.000144$	$[+0.000066, +0.000227]$	33/49	0.0017
Pooled explained variance	$+0.000141$	$[+0.000063, +0.000223]$	33/49	0.0015
Signed Pearson r	$+0.000211$	$[+0.000015, +0.000411]$	29/49	0.0649

CI excludes zero in $0/14$ series at $\tau = 1.30$ and $0/14$ at $\tau = 1.60$. Its components are supported unevenly: the early rise ($T8 - T2$) in $14/14$ series at $\tau = 1.30$ and $7/14$ at $\tau = 1.60$, the post-peak decline in $1/14$ and $8/14$, and an earlier fitted peak under the faster leak in $3/14$. We therefore describe the ratio trajectory and do not assert a parametric law for it; the bounded form that does win out of sample is the one fitted to the alignment curve of the continuous readout (§5.1).

Firing fraction per cell. Table 9 gives the fraction of layer-8 units that fire at each cell of the grid. At short unfolding almost no unit fires, which is the regime in which the continuous readout has not reached the plateau (§5.1) and in which the controlled contrast of Table 6 is negative.

B Limitations, broader impacts, and future work

B.1 Limitations

Cohort and stimulus scope. Our cohort is 49 native English speakers listening to a single audiobook (the English subset of Le Petit Prince (Li et al., 2022)). The 49 listeners give replication across listeners with the stimulus held constant, and no replication across stimuli or datasets; we cannot rule out that the plateau, the hybrid advantage, or the spatial heterogeneity partly reflect properties of this narrative.

Model-side integration. T and τ set how the converted network unfolds a fixed 128-token input. They are model-side computational variables and are not a proxy for cortical temporal receptive windows or for linguistic context length. The bounded plateau is a property of the model's encoding performance and is not a tested cortical claim.

Table 8: Controlled contrast (real minus matched random hybrid, pooled held-out R^2) per ROI: cell with the largest mean contrast, its value in units of 10^{-4}, and the FDR-positive cells across the 14-cell family of that ROI.

	$\tau = 1.30$			$\tau = 1.60$		
ROI	peak T	mean	FDR-positive T	peak T	mean	FDR-positive T
LanA	10	+4.78	10, 13, 16	6	+3.91	6, 8, 10
LH IFGorb	10	+4.02	none	8	+3.43	none
LH IFG	10	+7.69	10	8	+4.83	8
LH MFG	10	+1.97	none	6	+3.45	none
LH AntTemp	10	+6.07	10, 16	6	+3.76	6, 10, 16
LH MidTemp	10	+5.64	10, 16	6	+4.83	6, 8, 10
LH PostTemp	10	+0.46	none	8	+1.53	none
RH IFGorb	10	+5.30	10	6	+6.28	6, 16
RH IFG	10	+11.38	10, 13, 16	6	+7.75	6
RH MFG	10	+10.02	10	6	+9.02	6
RH AntTemp	10	+11.56	10, 13, 16	6	+9.91	6, 8, 10, 16
RH MidTemp	10	+12.72	10, 13	6	+11.05	6, 8, 10
RH PostTemp	10	+4.52	none	6	+3.33	none

Table 9: Fraction of layer-8 units that emit a spike, by cell (all nine runs).

τ	$T = 2$	4	6	8	10	13	16
1.30	0.000	0.001	0.008	0.154	0.713	0.883	0.946
1.60	0.001	0.039	0.692	0.892	0.958	0.990	0.998

Spike conversion approximation. Linear weights are bit-identical between the ANN and SNN arms, but at finite T the spike-converted forward pass is not numerically equal to the ANN's: trace_v is a continuous readout of the event code and only approximates the ANN activation (Eq. 1). Divergence outside the plateau therefore partly reflects conversion error. The two readouts are deterministically related (§3.2), so the hybrid result concerns representational format rather than an additional source of information.

Second substrate. The LAS replication changes the checkpoint and the conversion family together, so it shows that the temporal profile does not depend on SpikedAttention or on the matrix-attention parent without separating the two factors. Both parents are encoder-only BERT models. The hybrid contrast was repeated on LAS at one cell only, where it holds under held-out R^2 in $30/49$ listeners and not under signed r (Appendix A.9); the temporal grid of the control was not swept on LAS. At $T = 4$ the LAS substrate has a prediction-scale failure that correlation does not register.

fMRI temporal resolution and modality match. fMRI BOLD with $\mathrm{TR} = 2\,\mathrm{s}$ smooths over the millisecond-level dynamics of the spike-converted forward pass; we treat `spike_rate` as a within-TR sparse-recruitment axis, not as evidence about cortical spike timing. MEG, EEG, and ECoG operate at the same millisecond scale as the SNN substrate and are better matched to the spike-event framework. Our fMRI results should therefore be read as evidence about the slow, integrated component of language cortex; whether the substrate also aligns with the fast, event-locked component is a separate question.

Metric and framework scoping. The hybrid claim is supported under classic held-out R^2 together with $U_d = R^2_{cd} - R^2_c$; a Pearson^2 proxy on the same anchor cell inverts it (Appendix A.7), and the temporal profile of the controlled contrast differs under signed r (Appendix A.10). Generalisation to other metric or variance-partitioning choices is not implied.

Coverage of the matched control. The matched control was first evaluated at the anchor cell; Sweep (iv) extends it to the 14 cells of the temporal grid at layer 8, where it holds across the mid-to-long unfolding range and not uniformly; neighbouring layers were not swept. The control is capacity-

matched by construction, because candidate and control hybrids have the same dimensionality and pass through the same ridge pipeline. It does not test whether any 768-dimensional expansion of the continuous trace would produce a similar lift, and we record the absence of such capacity-only baselines as a limitation.

Spatial correspondence is descriptive. The match between the spatial amplitude ranking of the relative hybrid lift and the cortical temporal-receptive-window hierarchy (Hasson et al., 2008; Lerner et al., 2011; Honey et al., 2012) is descriptive and concerns the uncontrolled lift (Appendix A.10); we do not claim that cortex implements the specific spike code used in our substrate or that the lift causally tracks any specific linguistic operation.

Anchor-cell selection. The anchor cell $(T, \tau, L) = (10, 1.60, 8)$ was selected at the LanA group level and not per subject, to avoid double-dipping; per-subject selection coupled with within-subject inferential testing on the same data would induce selection bias. Group-level selection still uses the same dataset, and held-out selection would be stronger.

B.2 Future work

Several extensions follow from the present scoping. (i) *Cross-linguistic generalisation*: applying the same probe to the French and Mandarin runs of the Le Petit Prince corpus tests whether the bounded $T \times \tau$ plateau and the hybrid advantage are language-general or English-specific; because the narrative is shared, this varies cohort and language and not the stimulus. (ii) *Cross-stimulus generalisation*: replicating the controlled hybrid test on independent naturalistic-language fMRI corpora tests whether the results are stimulus-driven or substrate-driven. (iii) *Conversion families and backbones*: the temporal analysis now replicates on a second conversion family and parent (§5.1); a factorial design that separates backbone from conversion, a decoder-only substrate, and the hybrid test on further substrates are the next steps. (iv) *Time-resolved recordings*: ECoG or MEG with finer temporal resolution would let the within-TR sparse-recruitment interpretation of `spike_rate` be sharpened toward spike-timing hypotheses. (v) *Within-subject best-layer selection with proper cross-validation*: held-out within-subject layer selection (for example nested cross-validation across runs) could replace the current group-level anchor-cell choice and tighten parcel-specific claims.