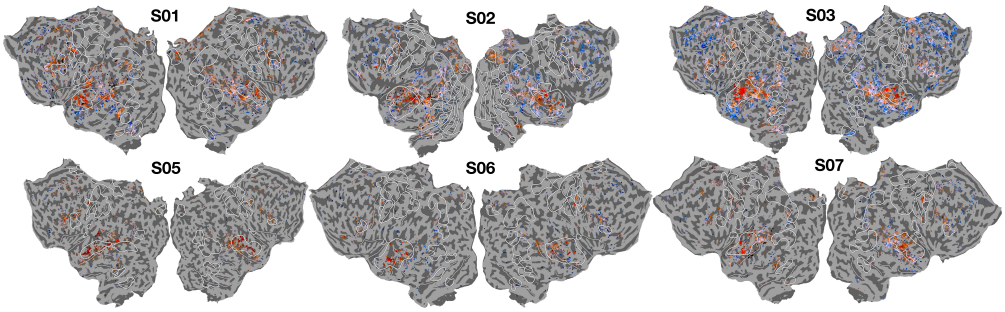

Figure 1: Timescales estimated in MT model (revised after bug fix). Colormap follows Fig. 3 in main text.

We thank the reviewers for their insights and suggestions. Due to limited space, we will address more minor comments in the camera-ready version should this paper be accepted. All references follow the main paper.

Common queries: Voxel timescale estimate T_v & negative β. As noted in supplementary section 1.3, there was a typo in the main text where Eq. 6 defines T_v. We use β_p^2, making T_v agnostic to $\beta' s$ sign. Further, the visualization suggested by reviewer 3 (plotting T_v vs. β_p) is shown in Supplementary Fig. 1B. This demonstrates the relationship between $|\beta|$ and T_v, serving as a proof of concept. **Literature as ground truth/Inaccuracy of δ-sum:** Firstly, section 4 demonstrates the inability of δ-sum to down-sample long timescale representations, causing these features to be highly correlated with word rate. We also find that AC voxels assigned *long* timescales by δ-sum are, in fact, well predicted by low-level models like word rate and acoustic spectrum, which change rapidly across time [4; Tang, Hamilton and Chang, 2017]. Taken together, this evidence strongly suggests that long T_v estimates by δ-sum are false and caused by the down-sampling confound. For precuneus and PFC, we find that δ-sum, in contrast, assigns shorter timescales than AC, going against the known language temporal hierarchy. **Leave-One-Out cross-validation for interpolation weights a.** The model learns a weight a_i on each word w_i to interpolate word activations W across time (Eq. 3). We solve for a in $a = \phi^{-1}W$ by ridge regression. The ridge coefficient is estimated by leaving one word out at a time and measuring accuracy of interpolating its activation from other words. **Encoding model fits rely on cross-validation.** To find regularization coefficients for β, we bootstrapped the regression procedure 50 times for each encoding model. In each bootstrap, a random set of 5000 TRs (125 blocks of 40 consecutive TRs) were removed from training set (26 stories) and used as validation data. Ridge coefficients were picked based on the validation set's prediction performance, averaged across bootstraps [4, 7]. Final model performance was computed on a separate test set (1 story).

R#1: Merits of timescale estimator over previous methods. To estimate timescale by manipulating context length, separate encoding models are first built for each CL. A voxel's CL preference is then computed as the center of mass of the *encoding performance curve* across different CLs. As discussed in the supplement, if the performance across CLs is similar (curve is flat, Supp. Fig 1A), the voxel has a *large* center-of-mass, artificially inflating the CL preference. This is the case with many AC voxels (Fig. 5). In comparison, our timescale estimation procedure is based on direct control of timescales in LMs, and predicts short-timescales in primary AC. **Brain areas integrate speech over 14s.** Prior work [2, 3, 8, 23] demonstrates through different experiments and methods that some brain regions integrate information over long timescales, on the order of several seconds. **Transformer LMs:** The main contribution of this paper is to use LMs with explicitly interpretable timescales to make detailed inferences about the brain. To the best of our knowledge, this is currently lacking in Transformer-based LMs. While we observe slightly better encoding performance with Transformer LMs (work under submission), to investigate the temporal hierarchy we are restricted to coarse CL preference estimates. **Cross-subject consistency:** The histograms in Figs. 4-5 compare different timescale estimation procedures across *all* 6 subjects. We found a bug in the colormap limits for subject S1, and show updated flatmaps in Fig. 1 here. The patterns are highly similar across subjects, as are the drawbacks of the other methods shown in the supplementary flatmaps. **Permutation test:** Block-wise permutation tests are entirely appropriate for assessing significance of model predictions in this setting, and account for temporal autocorrelation. An average of 6.8% of voxels in a subject are significant according to this test, demonstrating that non-permuted data doesn't always provide a better fit.

R#2: Restricting CL: The context length was restricted based on the back-propagation-through-time (BPTT) length in the baseline model upon which the interpretable LM was based (Merity et al. [21]).

R#3: Kernel choice: In practice, many kernels could be used for interpolation. However, the RBF kernel 1) generalizes to the δ-sum method when $\epsilon \to \infty$ (this was a typo in the main text) and 2) has a kernel width that can be directly linked to timescale. These properties are not exhibited by other commonly used alternatives, like polyharmonic spline kernels. **LM Performance:** Perplexity on WikiText2 test set (lower is better): 68.33 ± 0.12. Baseline LSTM (no timescale specification): 70.23 ± 0.24. These values are comparable to Merity et al. [21].