

Incremental Accumulation of Linguistic Context in Artificial and Biological Neural Networks

Supplementary materials

Table S1. Descriptive statistics of the data used in this study.

Narrative	# of participants	# of words	# of tokens	# of TRs
Lucy	14	1585	1767	351
Merlin	35	2134	2271	612
Pie man	73	954	1048	286
Tunnel	21	3309	3807	1013
Bronx	47	1356	1485	374
Sherlock	36	2509	2741	723
Not the fall	54	1561	1704	383
Milky way	17	1052	1155	282
Total	219 individuals	15260	15978	4314

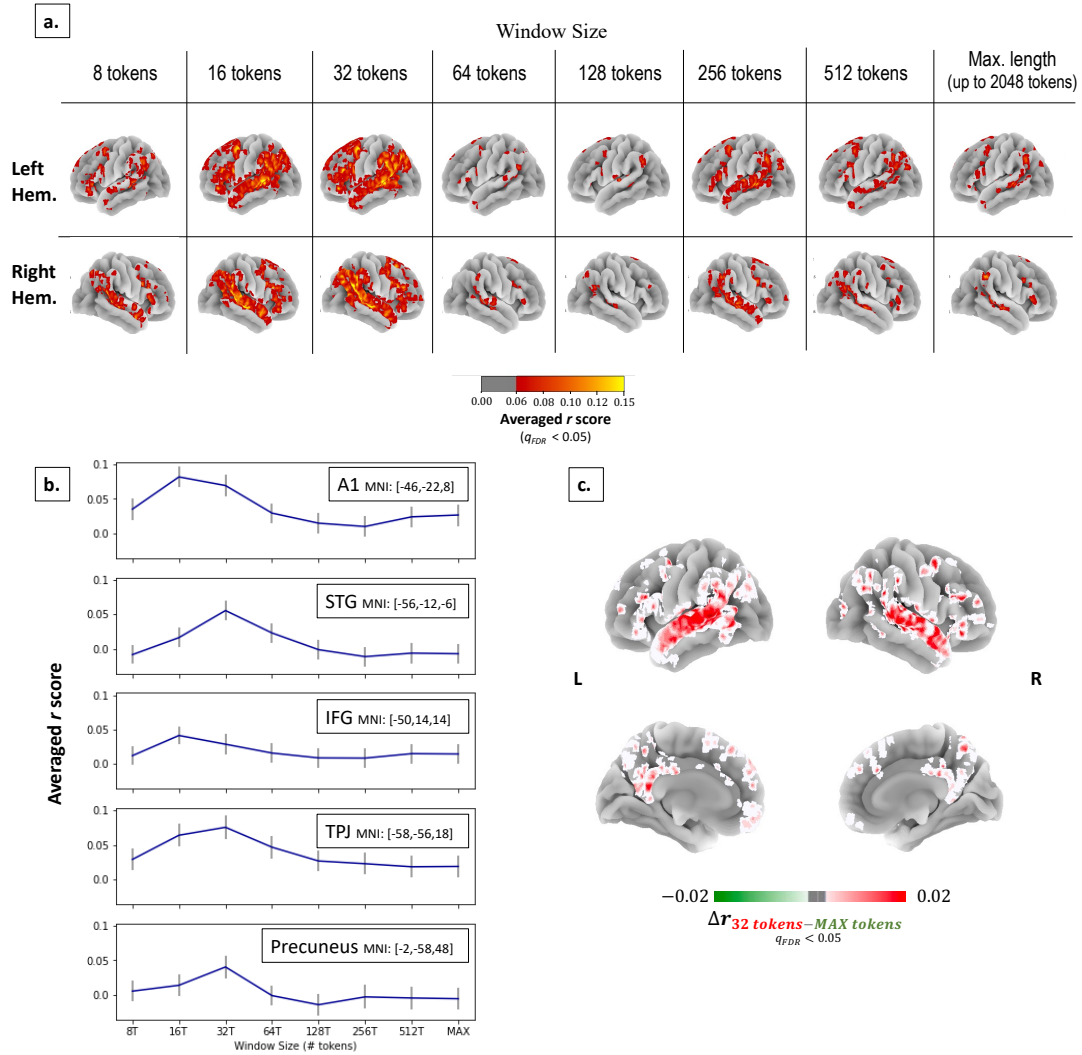


Figure S1. Replication of Figure 2 using the GPT-2 model. The effect of the size of the context window (in # of tokens) fed into the LLM on its ability to predict the neural signals. **a.** Cortical maps for different window sizes showing voxels where the neural encoder score (Pearson's r calculated between the predicted and the original signals, averaged across 297 fMRI scans) was statistically significant using a non-parametric Wilcoxon signed-rank test and FDR correction. **b.** Averaged r -scores by window size for five different voxels, each located in a different language-related region of interest (ROI) in the left hemisphere. Error bars represent the 95% confidence interval of the mean as calculated via 10,000-iterations bootstrap analysis. Anatomical locations provided as MNI coordinates. **c.** A cortical map showing ROIs that were predicted significantly better using a window size of 32 tokens than with a window size of up to 2048 tokens (red areas), or vice versa (green areas). Notes. A1 =

primary auditory cortex; STG = Superior Temporal Gyrus; IFG = Inferior Frontal Gyrus; TPJ = Temporoparietal Junction. Source data are provided as a Source Data file

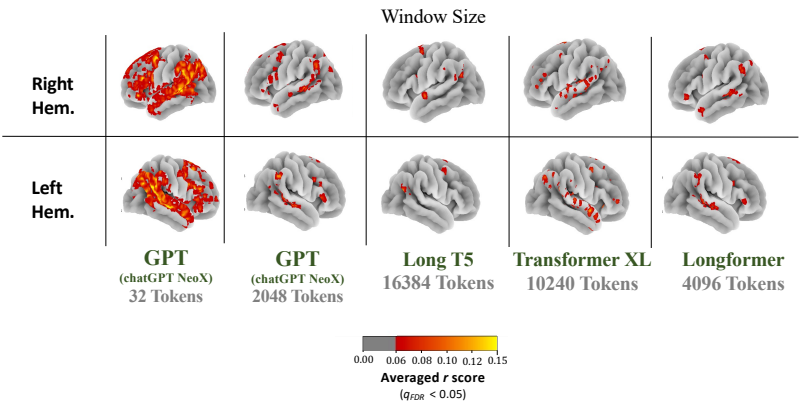


Figure S2. Cortical maps for different models/window sizes showing ROIs where the neural encoder score (Pearson's r calculated between the predicted and the original signals) was statistically significant. The pattern shown in Figure 2a is replicated here. When the window size of the transformer model exceeds 32 tokens, the neural encoder's performance decreases. This holds true even for models specifically designed for processing long texts.

a.

The Short-Term Context (last 32 tokens) Model

A window size of 32 tokens
(Short-term context)

... So I race back to the newsroom with my scoop and I find the editor, Jim Dwyer, who's a senior, and he will go on to win a Pulitzer **Prize**



b.

The Incremental Context Model

Ask the model to summarize the previous content, including the previous summary:

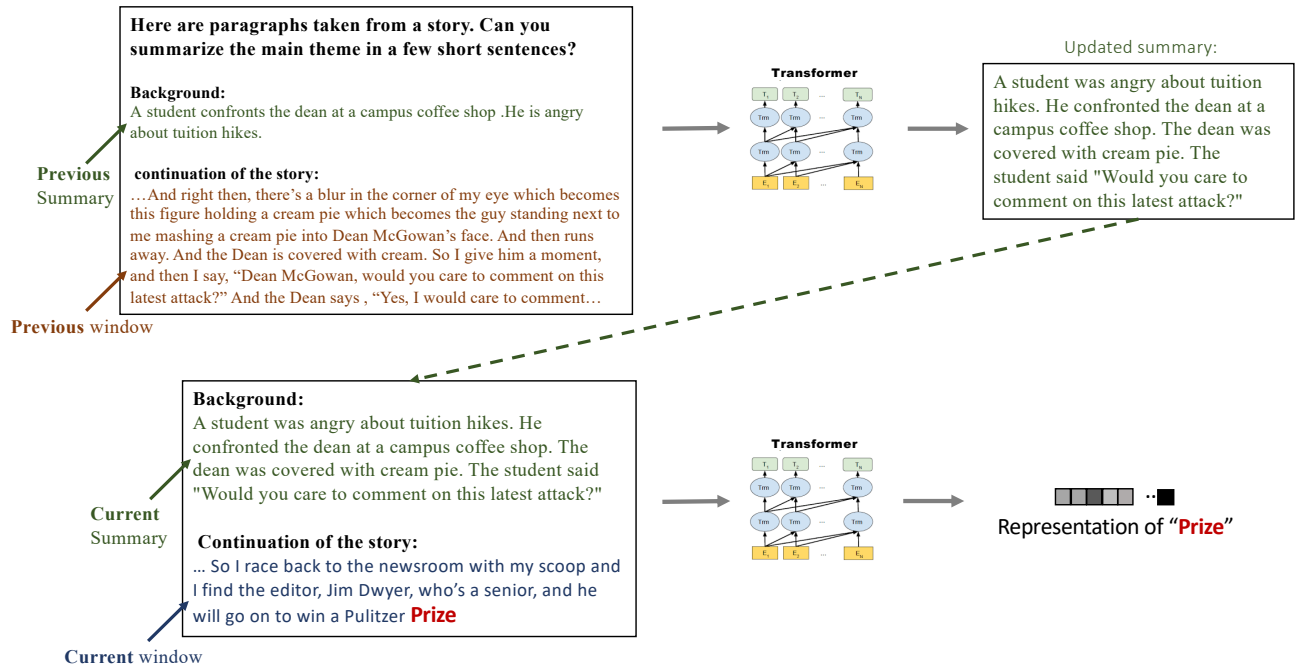


Figure S3. Schematic illustrations depict the process of extracting contextual embedding representations for a single token using both the baseline short-term context model (a) and the novel incremental context model (b) **a.** The short-term 32-tokens model. The word embedding representation of the token "Prize" is extracted by providing the LLM with that token, as well as the preceding 32 tokens. **b.** The long-term incremental context model. To extract the word embedding representation for the same token ("Prize"), the input to the model included both the short 32-token window (as in the 32-tokens model) and a concise natural-language summary generated by the model, which was based on information from the long-term context. This long-

term contextual summary is updated as the model progresses through the story. It is generated based on the text that appeared before the short 32-token window, as well as on the summary generated at the previous update step by the model.

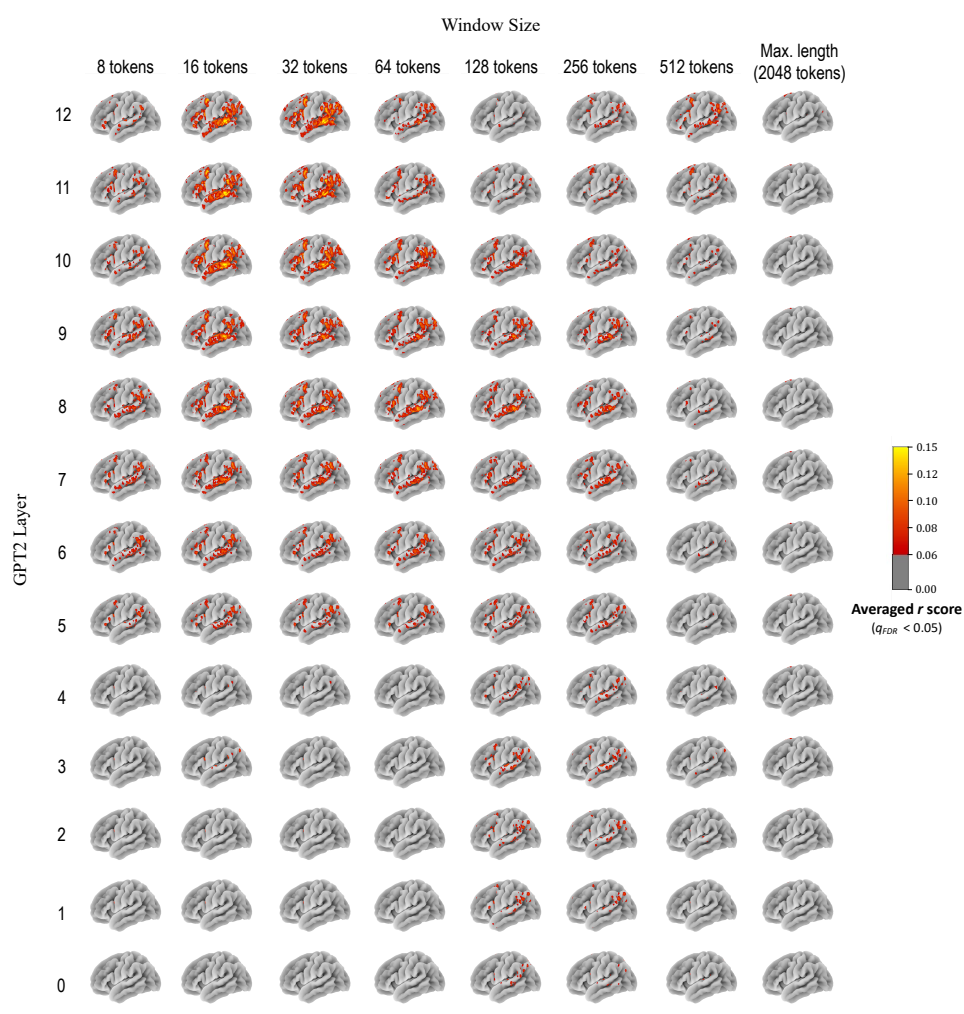


Figure S4. Replication of the analysis represented in Figure 2a using embedding representations extracted from different layers of the GPT2 model.