



HAL
open science

Deep language algorithms predict semantic comprehension from brain activity

Charlotte Caucheteux, Alexandre Gramfort, Jean-Rémi King

► **To cite this version:**

Charlotte Caucheteux, Alexandre Gramfort, Jean-Rémi King. Deep language algorithms predict semantic comprehension from brain activity. *Scientific Reports*, 2022, 10.1038/s41598-022-20460-9 . hal-03361426

HAL Id: hal-03361426

<https://hal.science/hal-03361426>

Submitted on 1 Oct 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

GPT-2's activations predict the degree of semantic comprehension in the human brain

Charlotte Caucheteux^{1,2,*}, Alexandre Gramfort², and Jean-Rémi King^{1,3}

¹Facebook AI Research, Paris, France; ²Université Paris-Saclay, Inria, CEA, Palaiseau, France; ³École normale supérieure, PSL University, CNRS, Paris, France

1 **Language transformers, like GPT-2, have demonstrated remarkable abilities to process text, and now constitute the backbone of deep translation, summarization and dialogue algorithms. However, whether these models encode information that relates to human comprehension remains controversial. Here, we show that the representations of GPT-2 not only map onto the brain responses to spoken stories, but also predict the extent to which subjects understand narratives. To this end, we analyze 101 subjects recorded with functional Magnetic Resonance Imaging while listening to 70 min of short stories. We then fit a linear model to predict brain activity from GPT-2's activations, and correlate this mapping with subjects' comprehension scores as assessed for each story. The results show that GPT-2's brain predictions significantly correlate with semantic comprehension. These effects are bilaterally distributed in the language network and peak with a correlation of $R=0.50$ in the angular gyrus. Overall, this study paves the way to model narrative comprehension in the brain through the lens of modern language algorithms.**

Neuroscience of language | Deep Neural Networks

1 **I**n less than two years, language transformers like GPT-2
2 have revolutionized the field of natural language processing
3 (NLP). These deep learning architectures are typically trained
4 on very large corpora to complete partially-masked texts, and
5 provide a one-fit-all solution to translation, summarization,
6 and question-answering tasks and algorithms (1).

7 Critically, their hidden representations have been shown to
8 – at least partially – correspond to those of the brain: single-
9 sample fMRI (2–4), MEG (2, 4), and intracranial responses to
10 spoken and written texts (3, 5) can be significantly predicted
11 from a linear combination of the hidden vectors generated
12 by these deep networks. Furthermore, the quality of these
13 predictions directly depends on the models' ability to complete
14 text (3, 4).

15 In spite of these achievements, strong doubts subsist on
16 whether language transformers actually encode meaningful
17 constructs (6). When asked to complete "I had \$20 and gave
18 \$10 away. Now, I thus have \$", GPT-2 predicts "20*". Simi-
19 lar trivial errors can be observed for geographical locations,
20 temporal ordering, pronoun attribution and causal reasoning.
21 These results have thus led some to argue that such "system
22 has no idea what it is talking about" (7). Thus, how the rep-
23 resentations of GPT-2 relate to a human-like understanding
24 remains largely unknown.

25 Here, we propose to evaluate how the similarity between the
26 brain and GPT-2 vary with semantic comprehension. Specifi-
27 cally, we first compare GPT-2's activations to the functional
28 Magnetic Resonance Imaging of 101 subjects listening to
29 70 min of seven short stories, and we quantify this similar-
30 ity with a "brain score" (\mathcal{M}) (8, 9). Second, we evaluate how

31 the brain scores systematically vary with semantic comprehen-
32 sion, as individually assessed by a questionnaire at the end of
33 each story.

34 **GPT-2's activations linearly map onto fMRI responses to spoken nar-**
35 **ratives.** To assess whether GPT-2 generates similar represen-
36 tations to those of the brain, we first evaluate, for each voxel,
37 subject and narrative independently, whether the fMRI re-
38 sponses can be predicted from a linear combination of GPT-2's
39 activations (Figure 1A). We summarize the precision of this
40 mapping with a brain score \mathcal{M} : i.e. the correlation between
41 the true fMRI responses and the fMRI responses linearly pre-
42 dicted, with cross-validation, from GPT-2's responses to the
43 same narratives (cf. Methods). To mitigate fMRI spatial
44 resolution and the necessity to correct each observation by
45 the number of statistical comparisons, we here report either 1)
46 the average brain scores across voxels or 2) the average score
47 within each region of interest ($n = 314$, following an automatic
48 subdivision of Destrieux atlas (10), cf. SI.1). Consistent with
49 previous findings (2, 4, 11, 12), these brain scores are signif-
50 icant over a distributed and bilateral cortical network, and
51 peak in middle- and superior-temporal gyri and sulci, as well
52 as in the supra-marginal and the infero-frontal cortex (2, 4, 11)
53 (Figure 1B).

54 By extracting GPT-2 activations from multiple layers (from
55 layer one to layer twelve), we confirm that middle layers best
56 map onto the brain (Figure 1C), as seen in previous studies
57 (2, 4, 11). For clarity, the following analyses focus on the
58 activations extracted from the *eighth* layer, i.e. GPT-2's most
59 "brain-like" layer (Figure 1B).

60 **GPT-2's brain predictions correlate with semantic comprehension.**
61 Does the linear mapping between GPT-2 and the brain reflect
62 a fortunate correspondence (4)? Or, on the contrary, does it
63 reflect similar representations of high-level semantics? To
64 address this issue, we correlate these brain scores to the level of
65 comprehension of the subjects, assessed for each subject-story
66 pair. On average across all voxels, this correlation reaches
67 $\mathcal{R} = 0.50$ ($p < 10^{-15}$, Figure 1D, as assessed across subject-
68 story pairs with the Pearson's test provided by SciPy). This
69 correlation is significant across a wide variety of the bilateral
70 temporal, parietal and prefrontal cortices typically linked to
71 language processing (Figure 1E). Together, these results sug-
72 gest that the shared representations between GPT-2 and the
73 brain reliably vary with semantic comprehension.

74 **Low-level processing only partially accounts for the correlation be-**
75 **tween comprehension and GPT-2's mapping** Low-level speech
76 representations typically vary with attention (13, 14), and
77 could thus, in turn, influence down-stream comprehension
78 processes. Consequently, one can legitimately wonder whether

*as assessed using Huggingface interface (<https://github.com/huggingface/transformers>) and GPT-2 pretrained model with temperature=0.

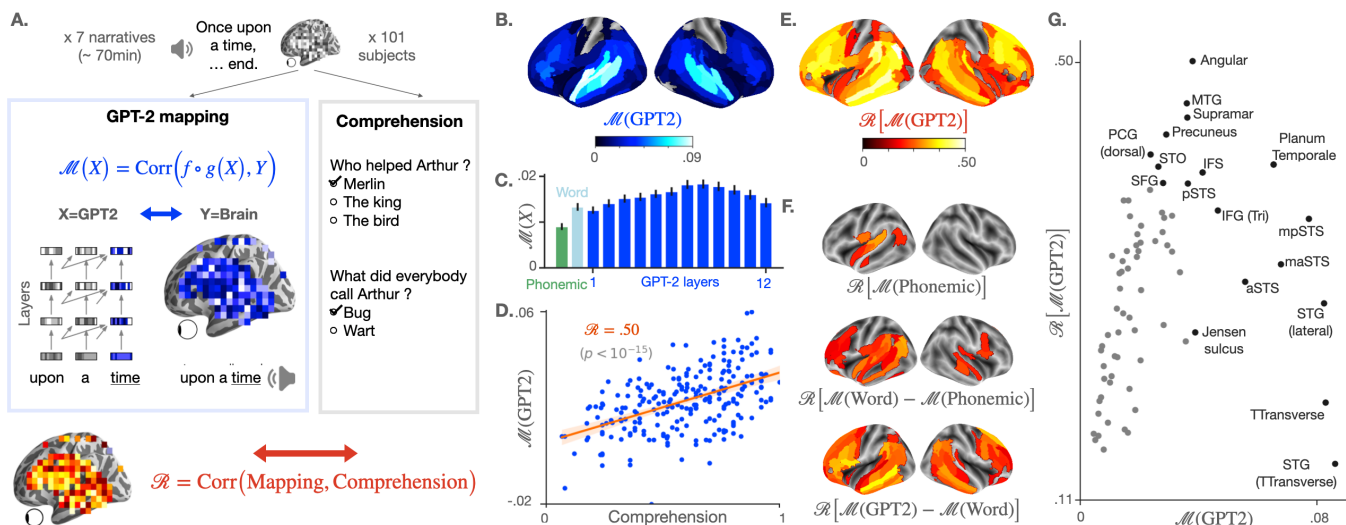


Fig. 1. A. 101 subjects listen to narratives (70 min of unique audio stimulus in total) while their brain signal is recorded using functional MRI. At the end of each story, a questionnaire is submitted to each subject to assess their understanding, and the answers are summarized into a comprehension score specific to each (narrative, subject) pair (grey box). In parallel (blue box on the left), we measure the mapping between the subject's brain activations and the activations of GPT-2, a deep network trained to predict a word given its past context, both elicited by the same narrative. To this end, a linear spatio-temporal model ($f \circ g$) is fitted to predict the brain activity of one voxel Y , given GPT-2 activations X as input. The degree of mapping, called "brain score" is defined for each voxel as the Pearson correlation between predicted and actual brain activity on held-out data (blue equation, cf. Methods). Finally, we test the correlation between the comprehension scores of the subjects and their corresponding brain scores using Pearson's correlation (red equation). A positive correlation means that the representations shared across the brain and GPT-2 are key for the subjects to understand a narrative. **B.** Brain scores (fMRI predictability) of the activations of the eighth layer of GPT-2. Scores are averaged across subjects, narratives, and voxels within brain regions (142 regions in each hemisphere, following a subdivision of Destrieux Atlas (10), cf. SI.1). Only significant regions are displayed, as assessed with a two-sided Wilcoxon test across (subject, narrative) pairs, testing whether the brain score is significantly different from zero (threshold: .05). **C.** Brain scores, averaged across fMRI voxels, for different activation spaces: phonological features (word rate, phoneme rate, phonemes, tone and stress, in green), the non-contextualized word embedding of GPT-2 ("Word", light blue) and the activations of the contextualized layers of GPT-2 (from layer one to layer twelve, in blue). The error bars refer to the standard error of the mean across (subject, narrative) pairs ($n=237$). **D.** Comprehension and GPT-2 brain scores, averaged across voxels, for each (subject, narrative) pair. In red, Pearson's correlation between the two (denoted \mathcal{R}), the corresponding regression line and the 95% confidence interval of the regression coefficient. **E.** Correlations (\mathcal{R}) between comprehension and brain scores over regions of interest. Brain scores are first averaged across voxels within brain regions (similar to B.), then correlated to the subjects' comprehension scores. Only significant correlations are displayed (threshold: .05). **F.** Correlation scores (\mathcal{R}) between comprehension and the subjects' brain mapping with phonological features ($\mathcal{M}(\text{Phonemic})$) (i), the share of the word-embedding mapping that is not accounted by phonological features $\mathcal{M}(\text{Word}) - \mathcal{M}(\text{Phonemic})$ (ii) and the share of the GPT-2 eighth layer's mapping not accounted by the word-embedding $\mathcal{M}(\text{GPT2}) - \mathcal{M}(\text{Word})$ (iii). **G.** Relationship between the average GPT-2-to-brain mapping (eighth layer) per region of interest (similar to B.), and the corresponding correlation with comprehension (\mathcal{R} , similar to D.). Only regions of the left hemisphere, significant in both B. and E. are displayed. In black, the top ten regions in terms of brain and correlation scores (cf. SI.1 for the acronyms). Significance in D, E and F is assessed with Pearson's p-value provided by SciPy[†]. In B, E and F, p-values are corrected for multiple comparison using a False Discovery Rate (Benjamin/Hochberg) over the 2×142 regions of interest.

79 the correlation between comprehension and GPT-2's brain
80 mapping is simply driven by variations in low-level auditory
81 processing. To address this issue, we evaluate the predictability
82 of fMRI given low-level phonological features: the word
83 rate, phoneme rate, phonemes, stress and tone of the narrative
84 (cf. Methods). The corresponding brain scores correlate with
85 the subjects' understanding ($\mathcal{R} = 0.17, p < 10^{-2}$) but less so
86 than the brain scores of GPT-2 ($\Delta\mathcal{R} = 0.32$). These low-level
87 correlations with comprehension peak in the left superior tem-
88 poral cortex (Figure 1F). Overall, this result suggests that the
89 link between comprehension and GPT-2's brain mapping may
90 be partially explained by – but not reduced to – the variations
91 of low-level auditory processing.

92 **The reliability of high-level representations best predict comprehen-**
93 **sion** Is the correlation between comprehension and GPT-2's
94 mapping driven by a *lexical* process and/or by an ability to
95 meaningfully *combine* words? To tackle this issue, we compare
96 the correlations obtained from GPT-2's word embedding (i.e.
97 layer 0) to those obtained from GPT-2's eighth layer, i.e. a
98 contextual embedding. On average across voxels, the correla-
99 tion with comprehension is 0.12 lower with GPT-2's word
100 embedding than with its contextual embedding. An analogous
101 analysis, comparing word embedding to phonological features

is displayed in 1F. Strictly lexical effects (word-embedding
versus phonological) peak in the superior-temporal lobe and
in pars triangularis. By contrast, higher-level effects (GPT-2
eighth layer *versus* word-embedding) peak in the superior-
frontal, posterior superior-temporal gyrus, in the precuneus
and in both the triangular and opercular parts of the inferior
frontal gyrus – a network typically associated with high-level
language comprehension (4, 15–19).

Comprehension effects are mainly driven by individuals' variability

The variability in comprehension scores could result from
exogeneous factors (e.g. some stories may be harder to com-
prehend than others for GPT-2) and/or from endogeneous
factors (e.g. some subjects may better understand specific
texts because of their prior knowledge). To address this issue,
we fit a linear mixed model to predict comprehension scores
given brain scores, specifying the narrative as a random effect
(cf. SI.1). The fixed effect of brain score (shared across nar-
ratives) is highly significant: $\beta = 0.04, p < 10^{-29}$, cf. SI.1).
However, the random effect (slope specific to each single nar-
rative) is not ($\beta < 10^{-2}, p > 0.11$). We also replicate the
main analysis (Figure 1D) within each single narrative: the
correlation with comprehension reaches 0.76 for the 'sherlock'
story and is above 0.40 for every story (cf. SI.1). Overall,

125 these analyses confirm that the link between GPT-2 and
126 semantic comprehension is mainly driven by subjects' individual
127 differences in their ability to make sense of the narratives.

128 **Discussion** Our analyses reveal a positive correlation between
129 semantic comprehension and the degree to which GPT-2 maps
130 onto brain responses to spoken narratives.

131 These results strengthen and complete prior work on the
132 brain bases of semantic comprehension. In particular, previous
133 studies have used inter-subject brain correlation to reveal the
134 brain regions associated with understanding (17). For exam-
135 ple, Lerner et al. recorded subjects' fMRI while they listened
136 to normal texts or texts scrambled at the word, sentence or
137 paragraph level, in order to parametrically manipulate their
138 level of comprehension (15). The corresponding fMRI signals
139 correlated across subjects in the primary and secondary audi-
140 tory areas even when the input was scrambled below the lexical
141 level. By contrast, fMRI signals also became correlated in the
142 bilateral infero-frontal and temporo-parietal cortex when the
143 scrambling was either not performed, or performed at the level
144 of sentences and paragraphs. Our results are consistent with
145 this hierarchical organization, and thus make an important
146 step towards the development of a cerebral model of narrative
147 comprehension.

148 The relationship between GPT-2's representations and hu-
149 man comprehension remains to be qualified. First, although
150 highly significant, our brain scores are relatively low (2, 9, 17).
151 This phenomenon likely results from a mixture of different
152 elements: i) we ran our analyses across *all* voxels to avoid
153 selection biases, which automatically reduces the average ef-
154 fect sizes and ii) we report the results without correcting for
155 a noise ceiling (cf. SI.1), as our pilot analyses suggest that
156 such noise-ceiling can greatly vary depending on how it is
157 implemented (i.e. fit from mean across subjects, from all or on
158 voxels etc). Second, the correlation between semantic compre-
159 hension and GPT-2's mapping is robust ($p < 10^{-15}$) but far
160 from perfect ($R = 0.50$). Such correlation thus indicates that
161 the modeling of brain responses with GPT-2 does not *fully*
162 account for the variation in comprehension. While this result
163 is expected (7), our study provides a promising framework to
164 evaluate the extent to which deep language models represent
165 and understand texts like we do.

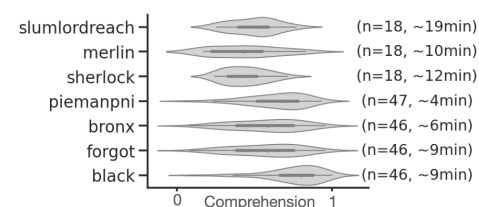
166 Finally, our results suggest that the neural bases of com-
167 prehension relate to the *high-level* representations of deep
168 language models. While the mapping of phonological fea-
169 tures and word embeddings do correlate with comprehension,
170 GPT-2's contextual embeddings provides brain maps that
171 more reliably predict comprehension (Figure 1F). The supe-
172 riority of contextual-embedding in predicting comprehension
173 suggests that i) GPT-2 encodes features supporting compre-
174 hension and ii) our finding are not solely driven by low- or
175 mid-level processing (13, 14). These elements remain solely
176 based on correlations, however. The factors that *causally* influ-
177 ence comprehension, ranging from prior knowledge, attention
178 and language complexity should be explicitly manipulated in
179 future work.

180 Overall, the present study strengthens and clarifies the simi-
181 larity between the brain and deep language models, repeatedly
182 observed in the past three years (2-4, 11, 20). Together, these
183 findings reinforce the relevance of deep language models in
184 unraveling the neural bases of narrative comprehension.

185 Materials and Methods

186
187 Our analyses rely on the "Narratives" dataset (21), composed of
188 the brain signals, recorded using fMRI, of 345 subjects listening to
189 27 narratives.

190 **Narratives and comprehension score** Among the 27 stories of the
191 dataset, we selected the seven stories for which subjects were asked
192 to answer a comprehension questionnaire at the end, and for which
193 the answers varied across subjects (more than ten different com-
194 prehension scores across subjects), resulting in 70 minutes of audio
195 stimuli in total, from four to 19 minutes per story (Figure 2). Ques-
196 tionnaires were either multiple-choice, fill-in-the blank, or open
197 questions (answered with free text) rated by humans (21). Here,
198 we used the comprehension score computed in the original dataset
199 which was either a proportion of correct answers or the sum of the
200 human ratings, scaled between 0 and 1 (21). It summarizes the
201 comprehension of one subject for one narrative (specific to each
202 (narrative, subject) pair).



203 **Fig. 2.** For each
204 of the seven narra-
205 tives: number of
206 subjects (n), distri-
207 bution of compre-
208 hension scores across
209 subjects and length
210 of the narrative.

211 **Brain activations** The brain activations of the 101 subject who
212 listened to the selected narratives were recorded using fMRI, as de-
213 scribed in (21). As suggested in the original paper, pairs of (subject,
214 narrative) were excluded because of noisy recordings, resulting in
215 237 pairs in total.

216 **GPT-2 activations** GPT-2 (1) is a high-performing neural language
217 model trained to predict a word given its previous context (it does
218 not have access to succeeding words), given millions of examples
219 (e.g Wikipedia texts). It consists of multiple Transformer modules
220 (twelve, each of them called "layer") stacked on a non-contextual
221 word embedding (a look-up table that outputs a single vector per
222 vocabulary word) (1). Each layer l can be seen as a nonlinear
223 system that takes a sequence of w words as input, and outputs
224 a contextual vector of dimension (w, d) , called the "activations"
225 of layer l ($d = 768$). Intermediate layers were shown to better
226 encode syntactic and semantic information than input and output
227 layers (22), and to better map onto brain activity (2, 4). Here, we
228 show that the *eighth* layer of GPT-2 best predicts brain activity
229 1C. We thus select the eighth layer of GPT-2 for our analyses.
230 Our conclusions remain unchanged with other intermediate-to-deep
231 layers of GPT-2 (from 6^{th} to 12^{th} layers).

232 In practice, the narratives' transcripts were formatted (replacing
233 special punctuation marks such as "-" and duplicated marks "?" by
234 dots), tokenized using GPT-2 tokenizer and input to the GPT-2
235 pretrained model provided by Huggingface ‡. The representation of
236 each token is computed separately using a context window a 1024.
237 For instance, to compute the representation of the third token of
238 the story, we input GPT-2 with the third, second and first token,
239 and then extract the activations corresponding to the third token.
240 To compute the representation of a token w_k at the end of the
241 story, GPT-2 is input with this token combined with the 1,023
242 preceding tokens. Then, we extract the activations corresponding
243 to w_k . The procedure results in a vector of activations of size (w, d)
244 with w the number of tokens in the story and d the dimensionality
245 of the model. There are fewer fMRI scans than words. Thus,
246 the activation vectors between successive fMRI measurements are
247 summed to obtain one vector of size d per measurement. To match
248 the fMRI measurements and the GPT-2 vectors over time, we used
249 the speech-to-text correspondences provided in the fMRI dataset
250 (21).

‡ <https://github.com/huggingface/transformers>

243 **Linear mapping between GPT-2 and the brain** For each (subject,
 244 narrative) pair, we measure the mapping between i) the fMRI
 245 activations elicited by the narrative and ii) the activations of GPT-2
 246 (layer nine) elicited by the same narrative. To this end, a linear
 247 spatiotemporal model is fitted on a train set to predict the fMRI
 248 scans given the GPT-2 activations as input. Then, the mapping is
 249 evaluated by computing the Pearson correlation between predicted
 250 and actual fMRI scans on a held out set I :

$$251 \quad \mathcal{M}^{(s,w)} : I \mapsto \mathcal{L} \left(f \circ g(X^{(w)})_{i \in I}, (Y_i^{(s,w)})_{i \in I} \right) \quad [1]$$

252 With $f \circ g$ the fitted estimator (g : temporal and f : spatial
 253 mappings), \mathcal{L} Pearson’s correlation, $X^{(w)}$ the activations of GPT-2
 254 and $Y^{(s,w)}$ the fMRI scans of subjects s , both elicited by the
 255 narrative w .

256 In practice, f is a ℓ_2 -penalized linear regression. We follow scikit-
 257 learn implementation[§] with ten possible regularization parameters
 258 log-spaced between 10^{-1} and 10^8 , one optimal parameter per voxel
 259 and leave-one-out cross-validation. g is a finite impulse response
 260 (FIR) model with 5 delays, where each delay sums the activations
 261 of GPT-2 input with the words presented between two TRs. For
 262 each (subject, narrative) pair, we split the corresponding fMRI time
 263 series into five contiguous chunks using scikit-learn cross-validation.
 264 The procedure is repeated across the five train (80% of the fMRI
 265 scans) and disjoint test folds (20% of the fMRI scans). Pearson
 266 correlations are averaged across folds to obtain a single score per
 267 (subject, narrative) pair. This score, denoted $\mathcal{M}(X)$ in Figure 1A,
 268 measures the mapping between the activations space X and the
 269 brain of one subject, elicited by one narrative.

270 **Phonological features** To account for low-level speech processing,
 271 we computed the alignment (Eq. (1)) between the fMRI brain recordings
 272 Y and phonological features X : the word rate (of dimension
 273 $d = 1$, the number of words per fMRI scan), the phoneme rate
 274 ($d = 1$, the number of phonemes per fMRI scan) and the concatenation
 275 of phonemes, stresses and tones of the words in the stimuli
 276 (categorical feature, $d = 117$). The latter features are provided in
 277 the original Narratives database (21), and computed using Gentle[¶]
 278 forced-alignment algorithm.

279 **Significance** Significance was either assessed by using either (i) a
 280 second-level Wilcoxon test (two-sided) across subject-narrative pairs,
 281 testing whether the mapping (one value per pair) was significantly
 282 different from zero (Figure 1B), or (ii) by using the first-level Pearson
 283 p-value provided by SciPy^{||} (Figure 1D-G). In Figure 1B, E, F, p-
 284 values were corrected for multiple comparison (2×142 ROIs) using
 285 False Discovery Rate (Benjamin/Hochberg)**.

286 References

287 1. Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Lan-
 288 guage Models are Unsupervised Multitask Learners. page 24, 2018.
 289 2. Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing
 290 (in machines) with natural language-processing (in the brain). *arXiv:1905.11833 [cs, q-bio]*,
 291 November 2019. arXiv: 1905.11833.
 292 3. Martin Schrimpf, Idan Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kan-
 293 wisher, Joshua Tenenbaum, and Evelina Fedorenko. Artificial Neural Networks Accurately
 294 Predict Language Processing in the Brain. *bioRxiv*, page 2020.06.26.174482, June 2020. .
 295 Publisher: Cold Spring Harbor Laboratory Section: New Results.
 296 4. Charlotte Caucheteux and Jean-Rémi King. Language processing in brains and deep neural
 297 networks: computational convergence and its limits. *bioRxiv*, page 2020.07.03.186288, July
 298 2020. . Publisher: Cold Spring Harbor Laboratory Section: New Results.
 299 5. Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey,
 300 Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan
 301 Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Fanda Lora, Adeen Flinker,
 302 Sasha Devore, Werner Doyle, Patricia Dugan, Daniel Friedman, Avinatan Hassidim, Michael
 303 Brenner, Yossi Matias, Ken A. Norman, Orrin Devinsky, and Uri Hasson. Thinking ahead:
 304 prediction in context as a keystone of language in humans and machines. *bioRxiv*, page
 305 2020.12.02.403477, January 2021. . Publisher: Cold Spring Harbor Laboratory Section:
 306 New Results.

307 6. Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 307
 Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint* 308
arXiv:1910.14599, 2019. 309
 7. Gary Marcus. Gpt-2 and the nature of intelligence. *The Gradient*, 2020. 310
 8. D. L. K. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo. 311
 Performance-optimized hierarchical models predict neural responses in higher visual cortex. 312
Proceedings of the National Academy of Sciences, 111(23):8619–8624, June 2014. ISSN 313
 0027-8424, 1091-6490. . 314
 9. Alexander G. Huth, Wendy A. de Heer, Thomas L. Griffiths, Frédéric E. Theunissen, and 315
 Jack L. Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. 316
Nature, 532(7600):453–458, April 2016. ISSN 0028-0836, 1476-4687. . 317
 10. Christophe Destrieux, Bruce Fischl, Anders Dale, and Eric Halgren. Automatic parcellation 318
 of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage*, 53 319
 (1):1–15, October 2010. ISSN 1053-8119. . 320
 11. Shailee Jain and Alexander G Huth. Incorporating Context into Language Encoding Models 321
 for fMRI. preprint, Neuroscience, May 2018. 322
 12. Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. 323
 Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, 324
 Daniel L. K. Yamins, and James J. DiCarlo. Brain-Score: Which Artificial Neural Network for 325
 Object Recognition is most Brain-Like? preprint, Neuroscience, September 2018. 326
 13. Nima Mesgarani and Edward F. Chang. Selective cortical representation of attended speaker 327
 in multi-talker speech perception. *Nature*, 485(7397):233–236, May 2012. ISSN 1476-4687. 328
 . Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7397 Primary_atype: Re- 329
 search Publisher: Nature Publishing Group Subject_term: Auditory system:Neuronal physi- 330
 ology:Perception Subject_term_id: auditory-system;neuronal-physiology;perception. 331
 14. Laurent Cohen, Philippine Salondy, Christophe Pallier, and Stanislas Dehaene. How does 332
 inattention affect written and spoken language processing? *Cortex*, 138:212–227, 2021. 333
 15. Y. Lerner, C. J. Honey, L. J. Silbert, and U. Hasson. Topographic Mapping of a Hierarchy 334
 of Temporal Receptive Windows Using a Narrated Story. *Journal of Neuroscience*, 31(8): 335
 2906–2915, February 2011. ISSN 0270-6474, 1529-2401. . 336
 16. C. Pallier, A.-D. Devauchelle, and S. Dehaene. Cortical representation of the constituent 337
 structure of sentences. *Proceedings of the National Academy of Sciences*, 108(6):2522– 338
 2527, February 2011. ISSN 0027-8424, 1091-6490. . 339
 17. Evelina Fedorenko, Terri Scott, Peter Brunner, William Coon, Brianna Pritchett, Gerwin 340
 Schalk, and Nancy Kanwisher. Neural correlate of the construction of sentence meaning. 341
Proceedings of the National Academy of Sciences of the United States of America, 113, 342
 September 2016. . 343
 18. Angela D. Friederici. The Brain Basis of Language Processing: From Structure to Function. 344
Physiological Reviews, 91(4):1357–1392, October 2011. ISSN 0031-9333, 1522-1210. . 345
 19. Gregory Hickok and David Poeppel. The cortical organization of speech processing. *Nature* 346
Reviews Neuroscience, 8(5):393–402, May 2007. ISSN 1471-0048. . Number: 5 Publisher: 347
 Nature Publishing Group. 348
 20. Jon Gauthier and Anna Ivanova. Does the brain represent words? An evaluation of brain 349
 decoding studies of language understanding. *arXiv:1806.00591 [cs]*, June 2018. arXiv: 350
 1806.00591. 351
 21. Samuel A. Nastase, Yun-Fei Liu, Hanna Hillman, Asieh Zadbood, Liat Hasenfratz, Negin 352
 Keshavarzian, Janice Chen, Christopher J. Honey, Yaara Yeshurun, Mor Regev, Mai Nguyen, 353
 Claire H. C. Chang, Christopher Baldassano, Olga Lositsky, Erez Simony, Michael A. Chow, 354
 Yuan Chang Leong, Paula P. Brooks, Emily Micciche, Gina Choe, Ariel Goldstein, Tamara 355
 Vandervel, Yaroslav O. Halchenko, Kenneth A. Norman, and Uri Hasson. Narratives: fMRI 356
 data for evaluating models of naturalistic language comprehension. preprint, Neuroscience, 357
 December 2020. 358
 22. Ganesh Jawahar, Benoît Sagot, and Djamel Seddah. What Does BERT Learn about the 359
 Structure of Language? In *Proceedings of the 57th Annual Meeting of the Association for* 360
Computational Linguistics, pages 3651–3657, Florence, Italy, 2019. Association for Compu- 361
 tational Linguistics. . 362

§ <https://scikit-learn.org/>

¶ <https://github.com/lowerquality/gentle>

|| <https://www.scipy.org/>

** <https://mne.tools/>

363 **Supporting Information (SI)**

364 **Brain parcellation.** In Figure 1B, E, and F, we used a subdivi-
 365 sion of the parcellation from Destrieux Atlas (10). Regions
 366 with more than 400 vertices were split into smaller regions (so
 367 that each regions contains less than 400 vertices). The original
 368 parcellation consists of 75 regions per hemisphere. Our custom
 369 parcellation consists in 142 regions per hemisphere.

370 In Figure 1G, we use the original parcellation for simplicity,
 371 and the following acronyms:

Acronym	Definition
STG / STS	Superior temporal gyrus / sulcus
aSTS	Anterior STS
maSTS	Mid-anterior STS
mpSTS	Mid-posterior STS
pSTS	Posterior STS
Angular / Supramar	Angular / Supramarginal inferior parietal gyrus
MTG / MTS	Medial temporal gyrus / sulcus
SFG / SFS	Superior frontal gyrus / sulcus
IFG / IFS	Inferior frontal gyrus / sulcus
Tri / Op	Pars triangularis / opercularis (IFG)
TTransverse	Temporal transverse sulcus
PCG	Posterior cingulate gyrus
STO	Temporo-occipital lateral sulcus

372 **Mixed-effect model.** Not all subjects listened to the same
 373 stories. To check that the \mathcal{R} scores (correlation between compre-
 374 hension and brain mapping) were not driven by the narratives
 375 and questionnaires' variability, a linear mixed-effect model was
 376 fit to predict the comprehension of a subject given its brain
 377 mapping scores, specifying the narrative as a random effect.
 378 More precisely, if $w_i \in \mathbb{R}$ corresponds to the mapping scores
 379 of the i^{th} subject that listened to the story w , and $C_{w_i} \in \mathbb{R}$
 380 refers to the comprehension scores, we estimate the fixed effect
 381 parameters $\beta \in \mathbb{R}$ and $\tilde{\eta} \in \mathbb{R}$ (shared across narratives), and
 382 the random effect parameter $\beta_w \in \mathbb{R}$ and $\eta_w \in \mathbb{R}$ (specific to
 383 the narrative w) such that:

$$C_{w_i} = (\tilde{\beta} + \beta_w) \times w_i + (\tilde{\eta} + \eta_w) + \epsilon_{w_i}$$

385 with ϵ_{w_i} a vector of i.i.d normal errors with mean 0 and vari-
 386 ance σ^2 . In practice, we use the statsmodels^{††} implementation
 387 of linear mixed-effect models. Significance of the coefficients
 388 were assessed with a t-test, as implemented in statsmodels.

389 **Replication across single narratives.** To further support that
 390 the \mathcal{R} were not driven by the narratives' variability, we repli-
 391 cate the analysis of Figure 1D within single narratives. In
 392 Figure 3, we show that correlation scores between brain scores
 393 and comprehension scores are positive for each of the seven
 394 narratives.

395 **Noise Ceiling Estimates.** fMRI recordings are inherently noisy.
 396 Thus, we estimate an upper bound of the best brain score that
 397 can be obtained given the level of noise in the Narrative dataset.
 398 To this end, for each (subject, narrative) pair, we linearly
 399 map the fMRI recordings, not with the GPT-2 activations,
 400 but with the average fMRI recordings of the other subjects
 401 who listened to that narrative. More precisely, we use the

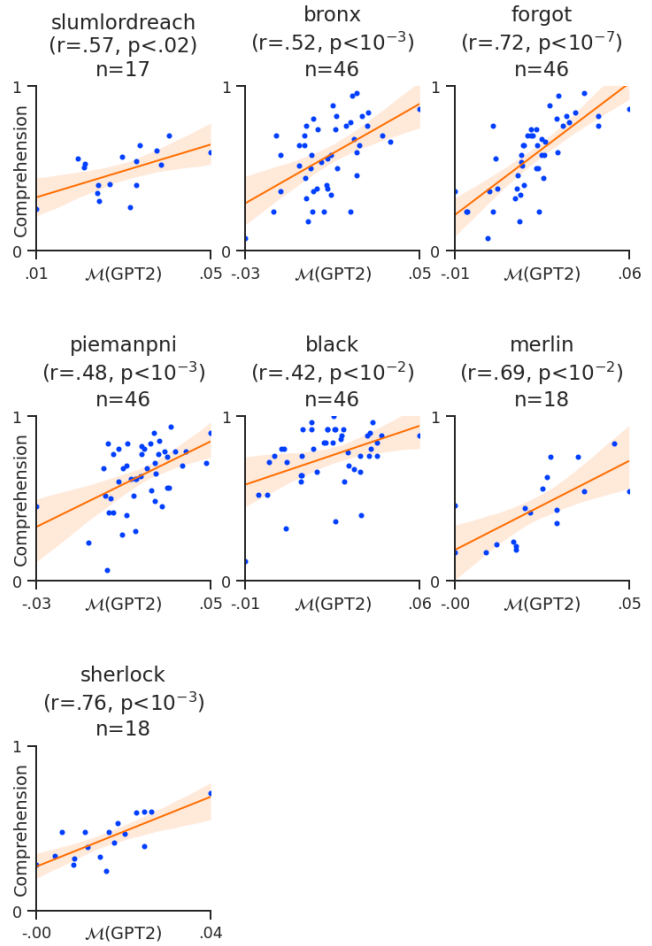


Fig. 3. Replication within single narratives. Same as Figure 1D for each single narrative.

exact same setting as in Eq. (1), but we predict $Y^{(s)}$, not
 from $g(X)$ (GPT-2's features after temporal alignment, of size
 $n_{\text{times}} \times n_{\text{dim}}$), but from the mean of the other subject's brains
 $\bar{Y} = \frac{1}{|\mathcal{S}|} \sum_{s' \neq s} Y^{(s')}$ (of size $n_{\text{times}} \times n_{\text{voxels}}$). This score is
 called the noise ceiling for the (subject, narrative) pair. The
 noise ceilings for each brain region are displayed in Figure 4,
 and correspond to upper bounds of the brain scores displayed
 in Figure 1B.

^{††} <https://www.statsmodels.org/>

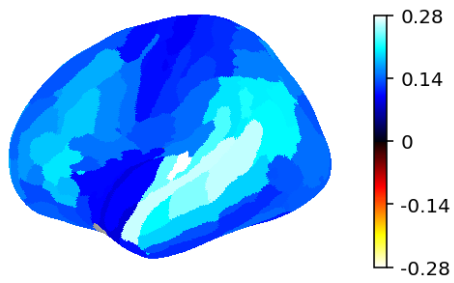


Fig. 4. Noise ceiling estimates. Noise ceilings averaged across subjects, narratives and voxels within each region of interest. They are upper bounds of the brain scores in Figure 1B.