OXFORD

Cerebral Cortex, 2023, **33**, 7830–7842

https://doi.org/10.1093/cercor/bhad082

Advance access publication date 20 March 2023

Original Article

Perspective changes in human listeners are aligned with the contextual transformation of the word embedding space

Refael Tikochinski[1,*], Ariel Goldstein[2], Yaara Yeshurun[3], Uri Hasson[2], Roi Reichart[1]

[1]Faculty of data and decision sciences, Technion—Israel Institute of Technology, Haifa 3200003, Israel,

[2]Department of Psychology and the Neuroscience Institute, Princeton University, Princeton, NJ 08544, USA,

[3]School of Psychological Sciences and Sagol School of Neuroscience, Tel Aviv University, Tel Aviv 69978, Israel

*Corresponding author: Faculty of data and decision sciences, Technion—Israel Institute of Technology, Haifa 320002, Israel.
Email: Rafyona@gmail.com

Word embedding representations have been shown to be effective in predicting human neural responses to lingual stimuli. While these representations are sensitive to the textual context, they lack the extratextual sources of context such as prior knowledge, thoughts, and beliefs, all of which constitute the listener's perspective. In this study, we propose conceptualizing the listeners' perspective as a source that induces changes in the embedding space. We relied on functional magnetic resonance imaging data collected by Yeshurun Y, Swanson S, Simony E, Chen J, Lazaridi C, Honey CJ, Hasson U. Same story, different story: the neural representation of interpretive frameworks. Psychol Sci. 2017:28(3):307–319, in which two groups of human listeners ($n = 40$) were listening to the same story but with different perspectives. Using a dedicated fine-tuning process, we created two modified versions of a word embedding space, corresponding to the two groups of listeners. We found that each transformed space was better fitted with neural responses of the corresponding group, and that the spatial distances between these spaces reflect both interpretational differences between the perspectives and the group-level neural differences. Together, our results demonstrate how aligning a continuous embedding space to a specific context can provide a novel way of modeling listeners' intrinsic perspectives.

Key words: computational modeling; neural encoding; listeners' perspective; contextual transformation.

Introduction

People process language in a very flexible and adaptable way. We are naturally able to encode the exact meaning of a word or phrase even in cases when it has several possible meanings, considering the context in which it appears. For example, the meaning of the word 'cold' varies between the contexts of cold weather, cold personality, and cold symptoms. One source of contextual information comes from the text itself, namely, from words or sentences that appear elsewhere in the text and shape the way the current word is interpreted. This phenomenon was extensively demonstrated in behavioral and neuroimaging studies (Federmeier et al. 2000; Van Berkum 2008; Nieuwland 2014; Van Berkum and Nieuwland 2019), and more recently also using deep language models (DLMs) that emerged from the natural language processing (NLP) field (Devlin et al. 2018; Peters et al. 2018; Radford et al. 2019). Another contextual information that shapes our language comprehension, and that was less investigated, is information that is not part of the text, but was obtained from other, external sources such as our thoughts, attitudes, and believes, all of which form our *perspective* regarding the text (Yeshurun et al. 2021). For example, while listening to a political debate, a listener may rely not just on the current and past statements of the speakers, but also on his own prior knowledge and political attitudes toward the topic. The current study suggests a novel computational framework to model this latter type of context, i.e. the human perspective, relying on the computational building-block of all recent state-of-the-art DLMs, the word embeddings representation.

In recent years, cognitive neuroscience researchers have started leveraging DLMs from NLP to better understand the neural mechanism of human language processing (Jain and Huth 2018; Schwartz et al. 2019; Caucheteux et al. 2021; Goldstein et al. 2022a). DLMs are massive artificial neural-network-based models of natural language capable of achieving human-level performances in many language tasks, including machine translation, text summarization, sentiment analysis, and more (Wu et al. 2016; Devlin et al. 2018, Peters et al. 2018; Aharoni et al. 2019). These models are trained using a massive amount of real-world texts, to predict the identity of missing words in a string of text (Devlin et al. 2018; Peters et al. 2018). As part of the neural computation, the DLM learns to represent each word via a multidimensional numerical vector, or a 'point' in a continuous space, known as the word embedding space. It turns out that the word embedding space effectively encodes many aspects of language, such as syntax, semantics, and pragmatics (Rogers et al. 2020). Interestingly, recent studies have demonstrated that word embedding vectors derived from such models can be used to predict neural responses in brain language areas during the processing of human language (Jain and Huth 2018; Schwartz et al. 2019; Caucheteux et al. 2021; Goldstein et al. 2022a). In addition, we recently found shared computational principles between the

Received: December 20, 2022. **Revised:** February 22, 2023. **Accepted:** February 23, 2023

ways the brain and DLMs process natural language (Goldstein et al. 2022b). These findings suggest that DLMs may serve as cognitive models for how humans process natural language.

The embedding space in DLMs is sensitive to the textual context. For example, the vector representation of the word 'cold' in the sentence 'you are cold as ice', will change as a function of other words in the text (e.g. whether the text discusses health issues or personality traits). But importantly, the embedding space is not affected by extratextual contextual sources that constitute the *listener's perspective* (i.e. previous thoughts, attitudes, and beliefs). To bridge this gap, we introduce a novel computational process by which we change the word embedding space to fit the language representation under a certain extratextual perspective. By this method, we can create multiple word embedding spaces so that each word embedding-space encodes language representations under a different human perspective. Our main assumption is that the variations we apply on the word embedding space can serve as a potential model for how humans change their language interpretation under different perspectives.

Multiple word embedding spaces in this study were created using fine-tuning—a common practice in deep learning (Hinton 2007). As mentioned above, DLMs are typically pre-trained using very large textual corpora (billions of words), sampled from a variety of textual domains and sources. This pre-training stage allows the model to learn how language is used across many natural contexts. Fine-tuning is a procedure used to adjust the embedding space to better fit to a specific narrower context (e.g. academic articles). In this study, we propose harnessing the fine-tuning technique to create a version of a DLM that fits a specific human perspective by designing an appropriate training setup and using a relevant dataset. Moreover, by designing multiple types of fine-tuning (i.e. using different datasets of different domains) we can create multiple versions of a DLM (and the corresponding word embedding space), each fits a different possible human perspective.

We test our idea using data from an functional magnetic resonance imaging (fMRI) experiment conducted by Yeshurun et al. (2017). In their experiment, two groups of participants ($n = 20$ for both groups) listened to the same audio recording of a short story by J.D. Salinger ('Pretty Mouth and Green My Eyes'). In the story, a husband loses track of his wife at a party and returns alone to their apartment in the city. Worried and anxious he calls his best friend, in the middle of the night, about the whereabouts of his wife. Next to the best friend, in bed, lies a mysterious woman whose identity is kept intentionally vague. Is she the wife, having an affair with the best friend (cheating context), or is she the friend's girlfriend, and the husband is unreasonably jealous as his friend implies (paranoia context)? Deciding between these two perspectives will have great consequences for interpreting the conversation. Before listening to the story, each experimental group was primed to adopt only one of these extratextual perspectives. The listener's perspective (cheating vs. paranoia) affected the neural responses to the story in areas with a long processing timescale, including the default mode network (DMN; Mars et al. 2012, Yeshurun et al. 2021), and frontal areas related to high-level language processing (Fletcher et al. 1995; Adolphs 2009; Mar 2011).

To implement our procedure in a way that simulates the two possible internal perspectives of Yeshurun et al.'s experiment, we created two alternative word embedding spaces by fine-tuning the DLM Bidirectional Encoder Representations from Transformers (BERT; Devlin et al. 2018) using either a dataset that focuses on 'cheating' stories, or a dataset that focuses on 'paranoia' stories. Our main aim is to test whether word embedding representations derived from the fine-tuned (cheating vs. paranoia) DLMs better fit the neural responses to Salinger's story in listeners with the matching perspective (cheating vs. paranoia). Such results demonstrate how aligning a continuous embedding space to a specific context can provide a novel way of modeling listeners' intrinsic perspectives.

Materials and methods

fMRI data

Participants, stimuli, and experimental design

The current study reanalyzes a previously published fMRI dataset (Yeshurun et al. 2017). The dataset consists of fMRI scans of 40 right-handed subjects assigned to one of the following experimental conditions: *Cheating* (10 females, 10 males, age: $M = 20.85$, *Standard deviation* $= 3.73$) or *Paranoia* (9 females, 11 males, age: $M = 21.45$, *Standard deviation* $= 3.42$).

The stimulus was a 11 min and 32 s record of a professional actor reading a short story of J.D. Salinger: 'Pretty Mouth and Green My Eyes.' The story describes a phone conversation between two friends, Arthur and Lee. Arthur has returned home after a party, and he lost track of his wife, Joanie. He is calling Lee to share his concerns over her whereabouts. Lee is at home, and a woman is lying on the bed next to him. The woman's identity is ambiguous—she may or may not be Joanie, Arthur's wife. Before listening to the story, participants were provided with a short introduction (~30 s) either specified that Arthur's wife is cheating on him with Lee (for the cheating condition), or that Arthur is paranoid and that his wife is not cheating on him (for the paranoia condition; Fig. 1). A story-comprehension questionnaire was administered immediately after the scan, and statistical analyses of the responses indicate that the context manipulation did affect the subject's interpretation of the story (see Yeshurun et al. 2017 for more details).

Preprocessing and voxel selection. The fMRI data were preprocessed by Yeshurun et al. (2017) and included the following steps: Motion correction, slice-time correction, linear-trend removal, high-pass filtering (two cycles per condition; ~0.003 Hz), spatial smoothing (Gaussian filter of 6-mm full width at half maximum (FWHM)), spatial transformation to 3-D Talairach space (Talairach and Tournoux 1988), and hemodynamic delay correction (based on the correlation between the audio envelope and the BOLD signal recorded from the A1).

To filter out stimulus-irrelevant voxels, we executed a voxel-wise inter-subject correlation analysis (ISC, Hasson et al. 2004) across all of the gray matter: For each voxel, we isolated each subject's time-course and correlated it with the averaged time-course of the remaining subjects. The voxel's ISC-score is calculated by averaging the correlation scores (after Fisher's Z transformation) obtained from repeating this process for all subjects. To assess how significantly each ISC-score is different from zero, we conducted a non-parametric permutation test by randomizing the phase of the signal (Simony et al. 2016) 1,000 times prior to ISC calculation and used the obtained null-distribution to estimate the *P*-value. We ran this procedure separately for each experimental group (cheating/paranoia) and selected for subsequent analyses only voxels that achieved a significant ISC-score ($P < 0.01$, corrected for multiple tests using FDR) in both groups. The process yielded a total of 11,783 'stimulus-locked' voxels.

Behavioral data

Besides neuroimaging data, we also used behavioral data collected by Yeshurun et al. (2017), which quantifies the effect

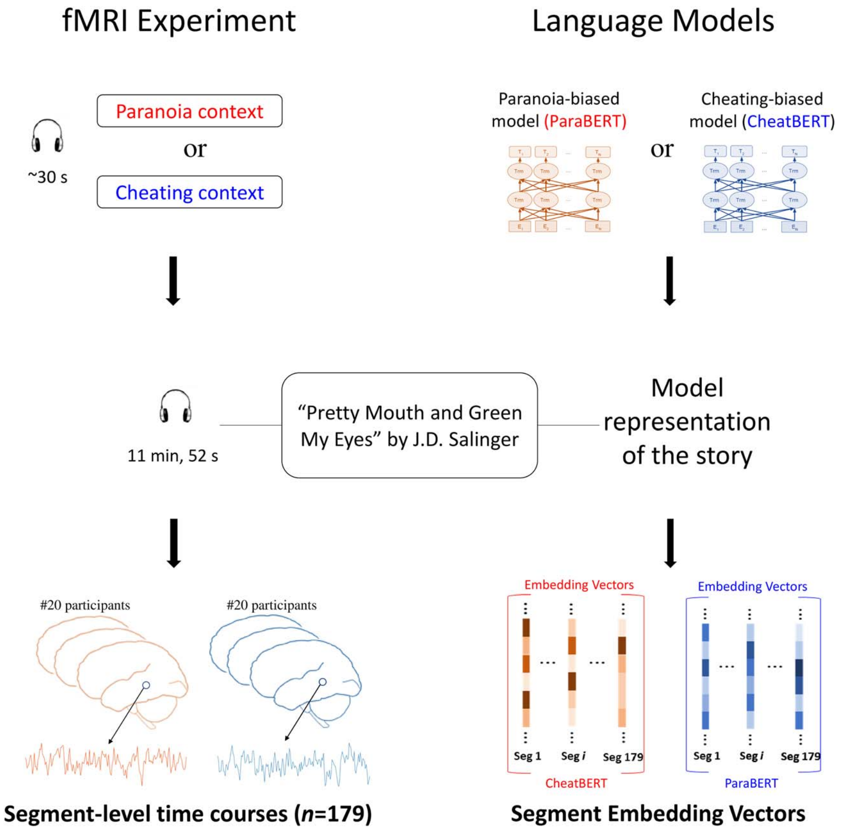
fMRI Experiment
Language Models
Paranoia context
or
Cheating context
~30 s
Paranoia-biased model (ParaBERT)
or
Cheating-biased model (CheatBERT)
"Pretty Mouth and Green My Eyes" by J.D. Salinger
11 min, 52 s
Model representation of the story
#20 participants
#20 participants
Embedding Vectors
Seg 1
Seg i
Seg 179
CheatBERT
Embedding Vectors
Seg 1
Seg i
Seg 179
ParaBERT
Segment-level time courses (n=179)
Segment Embedding Vectors

Fig. 1. An illustration of both the neural and the computational context-dependent representations of the same narrative. The left side represents the fMRI experiment by which the neural representations were acquired: 40 subjects were listening to an ambiguous short story that can be interpreted in two main contexts, cheating or paranoia. Half of the subjects were primed to the cheating context, and the other half were primed to the paranoia context. The right side illustrates the computational modeling of the experiment. Two context-dependent language models were created, each simulates a different context. We used each model to extract vector representations of the story (a.k.a. embedding vectors).

of context on participant's interpretation across the story. The text was divided into 179 segments (*Mean duration* = 3.77 s, *Standard deviation* = 2.39 s) by an independent expert annotator, and five independent raters were asked to rate how differently participants from different groups (cheating/paranoia) would interpret each segment. The raw scores (on a scale from 1 to 5) were first standardized to Z-scores for each rater, and then averaged across raters. The inter-rater reliability was high, as reflected by a Cronbach's α coefficient of 0.84 [See Yeshurun et al. (2017) for more details].

Computational modeling

We propose a novel method for computational modeling of context modulation in story interpretation. The main idea is to take a pre-trained language model and modify (fine-tune) it toward either the cheating or the paranoia contexts. Specifically, we used a well-accepted language model—BERT (Devlin et al. 2018)—as our initial, context-independent language model. This model was originally developed to be trained via a two-stage learning procedure: pre-training followed by fine-tuning and is considered the 'prototype' model of this type of learning. This is in contrast to other state-of-the-art language models, such as Generative Pre-trained Transformer (GPT; Radford et al. 2019) or T5 (Raffel et al. 2020), that are less compatible with the fine-tuning procedure. Accordingly, we next designed a fine-tuning process that creates two new context-dependent variants of BERT—CheatBERT (for the cheating context) and ParaBERT (for the paranoia context). We administered the fine-tuning process using dedicated datasets and classification tasks as described below.

Fine-tuning tasks and datasets. We defined two binary classification tasks: One to distinguish between cheating stories and no-cheating stories, and the other to distinguish between paranoia stories and no-paranoia stories. Our basic hypothesis is that fine-tuning BERT on these tasks (i.e. updating its parameters, which were set in the pre-training stage) would bias its internal representation of language toward the cheating (when using the first fine-tuning task) context or the paranoia context (when using the second fine-tuning task).

For these tasks, we collected 2,829 short stories (between 100 and 4,096 words; average number of words = 757.27, *Standard deviation* = 677.58; Fig. S2) concerning matters of relationships and romance. The stories were written by users of the Medium.com and Reddit.com websites. To locate relevant stories from Medium.com, the following website-tags were used: *marriage, relationships, romance, affairs, jealousy, monogamy, polygamy,* and *dating*. The stories from Reddit.com were extracted from the following subreddits: *askwoman, relationships, relationship_advice, romancestories, retroactive_jealousy, short_stories, sex,* and *teenagers*.

Each of the stories was manually tagged with one of the following three classes: Cheating ($n = 843$), Paranoia ($n = 1046$), and Other ($n = 940$). *There were two independent annotators, and stories for which there was disagreement were not used. The tagging was based on the reader's impression of the central theme emerging from the story, if it was mainly about cheating (affair), paranoia (jealousy), or none of them.* We used the stories labeled with *Cheating* and *Other* for the cheating vs. no-cheating classification (fine-tuning) task, and the stories labeled with *Paranoia* and *Other* for the paranoia vs. no-paranoia classification (fine-tuning) task.

Model architecture and the fine-tuning procedure. The same architecture was used for both the cheating- and the paranoia- classification models and is illustrated in Fig. S1. It consists of one classification head located on top of the original pre-trained version of the BERT encoder ('base' version, taken from the Huggingface repository). Since BERT's input is limited to a maximum of 512 tokens, and most of our stories are longer, a sliding window method (Pappagari et al. 2019) has been adopted. For each story, a fixed size window was 'moved' across the text, with an overlap between the windows (the size of the overlap as well as the size of the windows are both hyper-parameters of the model). Each textual window was fed into BERT, together with the standard prefix and suffix tokens (The "classification" token "CLS" and the "separation" token "SEP"), and the output vectors of all its tokens (apart from the CLS and the SEP tokens) were then averaged together, yielding a single window-representation vector.

Next, vectors from all the windows were fed into the classification head, which consists of one attention layer, and one fully connected nonlinear layer (with sigmoid activation) on top of it. The final output is a scalar ranging from 0 to 1, which represents the certainty of whether the story is about cheating or paranoia (1) or about other relationships related content (0).

The classifiers were trained to minimize the binary cross-entropy loss, using the batch gradient-descent algorithm (batch-size = 4). A total of 70% of the data was assigned for training, 15% for testing, and the remaining 15% for development and hyper-parameters calibration. The hyper-parameters included: window size {128,256,512}, overlap size {0,32,64,128}, the attention layer's dimension {64,256,512}, the number of neurons in the fully connected layer {32,64,128,256}, and the learning rate {1e − 3,1e − 4,1e − 5}. The training procedure ran epoch by epoch until no improvement was obtained in the model's predictions on the development data (The final number of epochs was 10 for CheatBERT, nine for ParaBERT and SpaceBERT; and eight for MedBERT, GunsBERT, and MideastBERT). Model performance was evaluated using a standard accuracy metric on the predicted scores (scores higher than or equal to 0.5 have been considered as 1, and lower than 0.5 as 0), which returns the proportion of the correct predictions.

Importantly, the original BERT's parameters from the top three (out of 12) layers were all updated during the training (the remaining parameters were maintained frozen due to a limited computational power).

Control models. Besides CheatBERT and ParaBERT, we created four additional BERT variants using the same fine-tuning procedure, but with different datasets and classification tasks. These models serve as control (baseline) models in our analyses. These variants are divided into two model pairs, where the members of each pair differ from each other in the specific context they model, but their contexts still refer to the same general theme (just as CheatBERT and ParaBERT both refer to the relationships and romance theme). The pairs were SpaceBERT—MedBERT (*Med* stands for medicine) and GunsBERT—MideastBERT. The models in the first pair are generally related to *science*, while those of the second pair are related to *politics*.

We used subsets of the publicly available 20*Newsgroup* dataset (http://qwone.com/∼jason/20Newsgroups/) for training (fine-tuning) these BERT-based classifiers: SpaceBERT was trained to distinguish texts tagged with *sci.space* (n = 987) from other science relevant texts (tagged with *sci*; n = 991); MedBERT was trained to distinguish texts tagged with *sci.med* (n = 990) from other science relevant texts (tagged with *sci*; n = 991); GunsBERT was trained to distinguish texts tagged with *politics.guns* (n = 780) from other politics relevant texts (tagged with *politics*; n = 685); Finally, MideastBERT was trained to distinguish texts tagged with *politics.mideasst* (n = 795) from other politics relevant texts (tagged with *politics*; n = 685).

Classifier evaluation. The first step in assessing the advantage of this procedure is to evaluate the performance of the difference classifiers. If this method is indeed effective, classifiers should achieve high accuracy scores when tested on the tasks they were trained on, but also, they should show a reduced accuracy in performance on tasks they were not trained on. All six classifiers resulted in a high accuracy score on the test data of their task, ranging from 0.80 to 0.95 (see Supplementary 2 and Table S1). Testing these classifiers on tasks that they were not trained on (e.g. testing the CheatBERT model on the paranoia classification task), indeed leads to a substantial performance drop. This implies that the models are all selectively specialized to the task they were trained on, and each one indeed captures a different and unique context.

Interestingly (and not surprisingly), as can be seen in Table S2, the performance declines are not uniform across all novel tasks, but rather they are affected by the global context they are sharing with each model. The performance of the models on novel tasks that belong to their global context (for example, CheatBERT and the paranoia classification task are sharing the same global context, which is romance and relationships), are better than their performance on novel tasks that do not relate to the global context (e.g. CheatBERT and the space-classification task).

Extracting neural and computational representations

The current research examines the relationships between neural and computational representations of the story. Representations were extracted segment-wise in accordance with the above mentioned (in the *Behavioral Data* section) textual segmentation of Salinger's story (n = 179 segments, Fig. 1). The BOLD signals of each stimulus-locked voxel (n = 11,783, see in the *fMRI Data* section) were 'down-sampled' from TR resolution (TR = 1.5 s) to a segment resolution by averaging all TRs within each segment (Mean number of TRs per segment = 2.51, *standard deviation* = 1.63).

We extracted segment-wise computational representations from each of our seven BERT variants [CheatBERT, ParaBERT, the original (pre-trained but not fine-tuned) BERT, and the four control models]. The extraction process was identical for all models since they all have the same architecture (12 attention-blocks stacked on top of each other). Each segment was fed to the model, together with a context of additional four segments—the two that preceded and the two that succeeded the relevant segment (whenever possible), as well as with the special tokens: CLS and SEP (These tokens represent the start/finish of the sentence). A vector representation of the segment was obtained by averaging only the embedding vectors (i.e. the output of layer 12 of the model) of the tokens, which belong to the relevant segment. This procedure yielded, for each model, a 179 (segment) by 768 (the BERT dimensionality) matrix (Fig. 1).

Since the dimensionality of the segment embedding vectors (768) is much higher than the number of samples in the data (179), we reduced the dimension of the vectors into 32 using principal component analysis (PCA; see also Goldstein et al. 2022a). PCA was calculated separately for each pair of models, and for the original BERT. Reducing to 32 dimensions provides a reasonable balance between the relatively low dimensionality and the relatively large fraction of the original variance preserved after the transformation (71% for CheatBERT-ParaBERT, 80% for SpaceBERT-MedBERT, 72% for GunsBERT-MideastBERT, and 68% for BERT). (The results below were also replicated using other dimensionalities, ranging from 16 to 75 dimensions.)

Semantic space analysis

The fine-tuning procedure has changed the internal parameters of BERT, and as a result, the word embedding space itself. The current analysis aims at investigating the word embedding spaces induced by CheatBERT and ParaBERT, and testing whether these changes are reasonably consistent with the manipulated perspectives (cheating/paranoia). To do that, we started by taking the 20 segments of the story that were rated by independent human-raters as those whose interpretation is most likely to change between cheating or paranoia perspectives (i.e. the top 20 segments from the *behavioral dissimilarity* scores). Next, we chose, a-priori, from each segment, the one or two keywords that best reflect the main semantic theme of the segment (there was a full agreement among the authors regarding each of the choices) and examined the extent to which the vector representations of the keywords were spatially changed between CheatBERT and ParaBERT. For each keyword, we calculated the cosine dissimilarity score between its vector representation, as obtained from CheatBERT, and its vector representation as obtained from ParaBERT. These cosine scores were then Z-normalized across all possible cosine scores, obtained from all the words of the story (1,876 words in total), so the keywords' distances (dissimilarity) scores can now be interpreted in relative to the distances obtained at random. We also do the same analysis, using all words in each section, to see that our method is robust for the selection of specific keywords. For comparison, we replicated these analyses using our control models (i.e. the vector representations obtained from the pairs: SpaceBERT—MedBERT, and MideastBERT—GunsBERT) as well.

Encoder-based context classification

We aim to show that our fine-tuned models (CheatBERT and ParaBERT) capture the information encoded in the brains of the participants, which belong to the corresponding group (CheatBERT for the *cheating* group and ParaBERT for the *paranoia* group). In other words, we hypothesize that the neural signal of each of the subjects would be better associated with the congruent model (e.g. the CheatBERT model with subjects from the cheating-condition group) than with the incongruent one (e.g. the ParaBERT model with subjects from the cheating condition and vice versa). Likewise, according to this hypothesis, we can use the models to *predict* the context in which a given subject listened to the story (cheating/paranoia), by correlating his/her neural signal with both CheatBERT and ParaBERT and checking which model is better correlated (Fig. 3a). By this logic, we formulated a voxel-wise classification task through that we can quantify the 'goodness of fit' of our models.

One way of measuring the level of association between DLMs and the brain is via *neural encoding* (Jain and Huth 2018; Goldstein et al. 2022a). A neural encoder is a voxel-wise linear regression model that takes as input word embedding vectors that were extracted from the DLM and predicts the corresponding level of the neural signal as recorded from a single voxel (usually a 1-dimensional scaler). From the neural-encoder model, we can extract a R^2_{adj}-score (adjusted coefficient of determination) and use it as an estimator for the level of association between the DLM and the neural signal. Our voxel-wise classifier exploits this neural-encoding scheme to measure the level of association between each model (CheatBERT and ParaBERT) and the subject's neural signal in a single voxel and predicts the context (*cheating* or *paranoia*) according to the model that achieved the highest R^2_{adj} score. The full process is described below.

For each stimulus-locked voxel (11,783 voxels with reliable ISC, see in the *fMRI Data* section), the classifier iterates over all subjects' brains ($n = 40$) and predicts the context (cheating/paranoia) as follows: First, we fit two linear regression models to predict the neural time-course from the vector representations of the story (a.k.a 'neural encoder,' see Jain and Huth, 2018; Goldstein et al. 2022a). The linear regression takes as input the vector representation of a segment (a 32-dimensional vector, see Section: '*Extracting neural and computational representations*') and predicts the averaged BOLD signal corresponding to that segment. One model uses vectors extracted from CheatBERT, and the other uses the ParaBERT's vectors. Then, we calculate the R^2_{adj} score (adjusted coefficient of determination) from each model and classify the context in accordance with the best R^2_{adj} score. Namely, if the CheatBERT's R^2_{adj} score is higher than the ParaBERT's R^2_{adj} score, we will classify that brain as *cheating*, and vice versa (Fig. 3a). We evaluate the classifier by calculating the accuracy rate of its prediction (the number of correct predictions divided by 40). This procedure provides a single accuracy score for each voxel, which quantifies the extent to which our models fit the neural signal in different brain areas.

We repeated the same analysis using other pairs of control models, for the purpose of comparison. The alternative pairs were: CheatBERT vs. BERT, BERT vs. ParaBERT, GunsBERT vs. MideastBERT, and MedBERT vs. SpaceBERT. The significance testing of this analysis is described in the *Statistical analysis* section below.

Distance analysis

The neural modulation caused by the context is not uniform, but varies throughout the story: there are parts of the story that cause more substantial neural differences between brains compared to other parts of the story (Yeshurun et al. 2017). We wanted to test whether this dynamic is also encoded in our fine-tuned models. To test this, we calculated the distance between each pair of segment embedding vectors (extracted from CheatBERT and ParaBERT) using the cosine distance (which is equal to 1 minus the cosine similarity of the vectors). This process yielded a 179-dimensional distance vector (corresponding to the 179 segments of the story, Fig. 3b). Likewise, we calculated the neural differences (distance) between the average brain activity of the *cheating* group and the average brain activity of the *paranoia* group. This is done by taking the absolute values of the differences between the averaged brain activities of each segment in every stimulus-locked voxel. This process yielded a single 11,783 (voxels) by 179 (segments) matrix of neural distance scores. Finally, we calculated the correlation between each voxel's neural distance and the model's distance vector using Pearson's r (Fig. 3b).

In addition, we analyzed the correlation between the models' distance vector and the differences in human interpretation of the story (the behavioral measurement, see in the *Behavioral data* section). The analyses were repeated using other distance vectors extracted from the following pairs of control models: CheatBERT vs. BERT, BERT vs. ParaBERT, GunsBERT vs. MideastBERT, and MedBERT vs. SpaceBERT. The significance tests of these analyses are detailed below in the *Statistical analysis* section.

Statistical analysis

All analyses were tested for statistical significance using non-parametric permutation tests. In the classification analysis (the

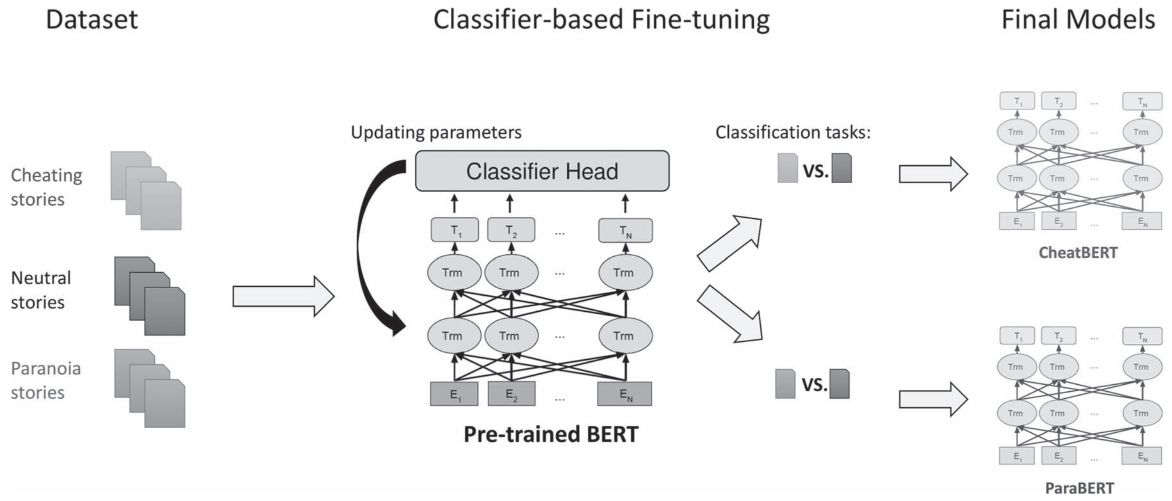

Fig. 2. The classifier-based fine-tuning process. We aim to create two new context-dependent language models by changing the parameters of an existing pre-trained model (BERT) to point toward the cheating or the paranoia context. First, we collected ~3,000 short stories, tagged as cheating, paranoia, or neutral stories. Then, we trained (fine-tuned) a BERT-based classifier (the BERT model with a classification head; Fig. S1 and Table S1) on either the cheating vs. no-cheating or the paranoia vs. no-paranoia classification task, updating the parameters of both the pre-trained BERT model and of its classification head. Finally, we removed the classifier head and were left with the new, cheating- (or paranoia-) induced variant of BERT: CheatBERT (or ParaBERT).

Encoder-based context classification section) we tested the significance of the accuracy scores by creating an estimated null-distribution using 1,000 permutations of the data. In every permutation step we shuffled the labels (i.e. cheating/paranoia) of the participants, ran the classification analysis on that randomized data, and saved the accuracy scores. The procedure returns a different 1,000-sized distribution for each voxel (for a total of 11,783 voxels). To account for multiple hypothesis testing, we calculated the *P*-values of the observed (real) accuracy scores using family wise error rate (FWER; Nichols and Hayasaka 2003) estimation: We combined the 11,783 distributions into a single, 1,000-sized distribution by taking only the maximum value (i.e. the best voxel's accuracy score) from each permutation step. Next, we calculated the *p*(FWER) score from the obtained max-values null-distribution using the following formula:

$$p(\text{FWER}) = (k + 1)/1{,}000,$$

where *k* is the number of max-values larger than the real value. We considered a voxel's score as significant if its *p*(FWER) was smaller than 0.05.

The same procedure was applied for the remaining analyses and the only difference was regarding the way we permuted the data. In the distance analysis (the *Distance analysis* section) we calculated the *p*(FWER) of each Pearson's *r* score using the max-values null-distribution obtained from 1,000 permutations, as above, but the data was permuted using randomized phase-shuffling. This method randomizes the signal while maintaining the exact mean and autocorrelation as the original signal (Simony et al. 2016). We implemented this shuffling method by applying a fast-Fourier transformation on the original signal, randomizing only the phase component of the signal, and then applying an inverse fast-Fourier transformation using the original frequency magnitudes and the randomized phases. The shuffling was performed only on the neural signals. In the models-behavior correlation analysis we shuffled only the behavioral signal.

Code and data

The dataset we collected for this study, as well as the code for our analyses are available for free in *https://github.com/RefaelTikochinski/Modeling-perspective-changes-in-human-listeners*.

Results

Same story, different word embedding representation

Using fine-tuning, we induced changes in the word embeddings space of the pre-trained model BERT (This DLM was specifically designed for fine-tuning, unlike other state-of-the-art models.) (Devlin et al. 2018), creating two alternative word embedding spaces, CheatBERT and ParaBERT, one for each of the two possible listener's perspectives in listening to the J.D. Salinger's story, cheating or paranoia. CheatBERT was created by fine-tuning BERT to distinguish between cheating and no-cheating stories and ParaBERT was created by fine-tuning BERT to distinguish between paranoia and no-paranoia stories (See Materials and methods; Figs. 1 and 2). Next, we extracted embedding representations of Salinger's story using both CheatBERT and ParaBERT. This process yielded two different sets of vectors, each represents the exact same story, but with a different perspective (Fig. 1). In addition, for comparison, we also created another four 'control' word embedding spaces using the same fine-tuning procedure but with completely different datasets and topics (SpaceBERT, MedBERT, MideastBERT, and GunsBERT; see Materials and methods).

Before testing our procedure on the fMRI data, we conducted a sort of 'manipulation-check' to test whether the various word embeddings spaces capture changes in the semantic interpretation across the two perspectives (cheating/paranoia). To do that, we started by taking the 20 segments of the story that were rated by independent human-raters as those whose interpretation is most likely to change between the cheating and paranoia perspectives (i.e. the top 20 segments from the *behavioral dissimilarity* scores; see Materials and methods). Next, in each segment, we focused on the change in the embedding representation across the two word embedding spaces (Z-normalized cosine distance, see Materials and methods) for keywords that drive the different interpretations across perspectives. We also performed the same analysis using all words in each section, to see that our method is robust for the selection of specific keywords. For example, in the top section, Arthur [the husband] asks Lee [the friend]: *'did you happen to notice when Joanie* [the wife] *was leaving?'* Lee replies: *'No, I didn't, Arthur'*. The word 'No' is the keyword of that segment, as

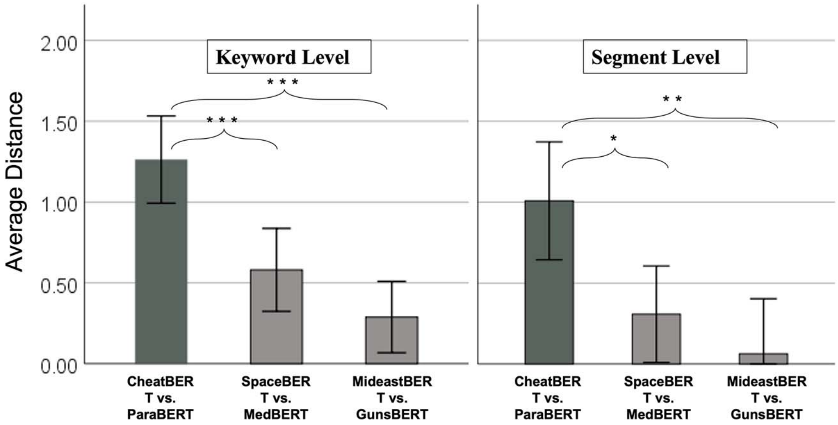

Fig. 3. The extent of the change between the vector representations of the different models in the top 20 relevant segments (i.e. segments that were rated by independent human raters as those whose interpretation is most likely to change between cheating or paranoia perspectives). The figure presents the averaged Z-normalized cosine-distance score of either selected keywords (left panel) or the entire segments (right panel), as calculated between CheatBERT's and ParaBERT's representations, compared to the scores obtained from the control model paires. $^*P < 0.05$, $^{**}P < 0.005$, $^{***}P < 0.00001$.

its hidden intention subtly changes under the two perspectives. According to the cheating perspective, the 'No' hints of a lie, since Arthur's wife is actually in bed with him at that moment. In contrast, from the paranoia perspective, the 'No' reflects the accurate state of affairs given that the woman in Lee's bed is not Joanie but his legitimate girlfriend.

Indeed, the Z-normalized distance between the CheatBERT's and the ParaBERT's vector representations of the word 'No' was 1.33, that is, 1.33 standard-deviation units above the average distance score of all the words in the story. In contrast, the Z-normalized scores obtained from the control models were lower and much closer to the averaged distance score of all the words in the story: 0.18 and 0.80 for SpaceBERT—MedBERT and Mideast-BERT—GunsBERT, respectively. Similar results were obtained for 26 keywords across the 20 segments (Table S3; this table also contains the bottom 20 segments for comparison). To calculate significance, we calculated the average Z-normalized distance between the CheatBERT's and the ParaBERT's word embeddings across all 20 segments, and demonstrated how it was significantly higher than the distance for the same keywords across the control fine-tuned models (Fig. 3A). While slightly lower, the same pattern of results holds when we include all words in the segment (Fig. 3B). Together, our results imply that our fine-tuning procedure modifies the relevant areas of the word embedding space in a way that properly reflects the change in perspective of the human listeners.

Variations on the word embedding space fit contextual modulation in the brain

Creating two alternative word embedding spaces (CheatBERT and ParaBERT) allows us to model perspective-based unique neural responses of our listeners. First, in line with previous studies, the word embeddings representations of Salinger's story, as extracted from either CheatBERT or ParaBERT, were highly effective in predicting the averaged neural signal of both group of listeners, using a standard voxel-wised neural encoder (Jain and Huth 2018; Schwartz et al. 2019; Caucheteux et al. 2021; see Materials and methods). For the cheating group, the CheatBERT model significantly predicted 8,226 voxels (69.8% of all stimulus-locked voxels, $0.26 < R^2_{adj} < 0.59$, *mean* $R^2_{adj} = 0.37$) and the ParaBERT model significantly predicted 5,998 voxels (50.09%, $0.26 < R^2_{adj} < 0.58$, *mean* $R^2_{adj} = 0.35$; Figs. 5A and S4). For the Paranoia group, the ParaBERT model significantly predicted 6,692 voxels (We attribute the smaller number of predicted voxels in the paranoia group to the fact that in this group the between-subjects variance is much larger than in the cheating group, as reflected in a relatively lower ISC-score [ISC; (See Fig. S3 and Materials and methods)]. Explaining this observation is beyond the scope of this study and is kept for future work.) (56.7%, $0.24 < R^2_{adj} < 0.60$, *mean* $R^2_{adj} = 0.35$) and the CheatBERT model predicted 3,774 voxels (32%, $0.30 < R^2_{adj} < 0.57$, *mean* $R^2_{adj} = 0.38$; Figs. 5a and S4. For a comparison of fine-tuned to original BERT, see Supplementary 3 and Fig. S4). Importantly, we detected a subset of voxels in which the neural signal was better predicted by the congruent models (i.e. CheatBERT for the cheating groups and ParaBERT for the paranoia group) than by the incongruent models (e.g. CheatBERT for the paranoia group). Specifically, we applied our novel encoder-based classification analysis (Fig. 4A; see Materials and methods) that its accuracy score reflects the proportion of listeners whose neural activity was better predicted with the congruent model. We perform the analysis on a voxel-by-voxel level, among all the stimulus-locked voxels ($n = 11,783$; Fig. S3). The classifier accuracy rate was above the chance level in 921 voxels, ranging from 62.5% to 85% ($P < 0.05$, family wise error corrected). These voxels encompass brain regions that resemble those typically found in the DMN [Bilateral temporoparietal junction (TPJ), middle frontal gyrus, and Precuneus; see Yeshurun et al. 2021], as well as in the right ventrolateral prefrontal cortex (vLPFC), bilateral superior temporal gyrus (STG), and middle temporal gyrus (MTG) (Fig. 5B and Table 1). Hence, neural responses in these brain areas that are unique to each group of listeners (cheating and paranoia group), are significantly associated with the unique information encoded in the corresponding fine-tuned model (CheatBERT and ParaBERT).

The original pre-trained BERT model or the BERT models that are fine-tuned in unrelated contexts were not beneficial in classifying the listener's perspective. Running the same classification procedure but replacing ParaBERT with BERT yielded only 25 significant voxels (17 voxels from the vLPFC and eight voxels from the cuneus, max scores: 70% and 65%, respectively), and replacing CheatBERT with BERT yielded zero significant voxels. Likewise, we found no significant voxels when replacing the ParaBERT and CheatBERT models with the other control-models pairs, MedBERT-SpaceBERT, and GunsBERT-MideastBERT.

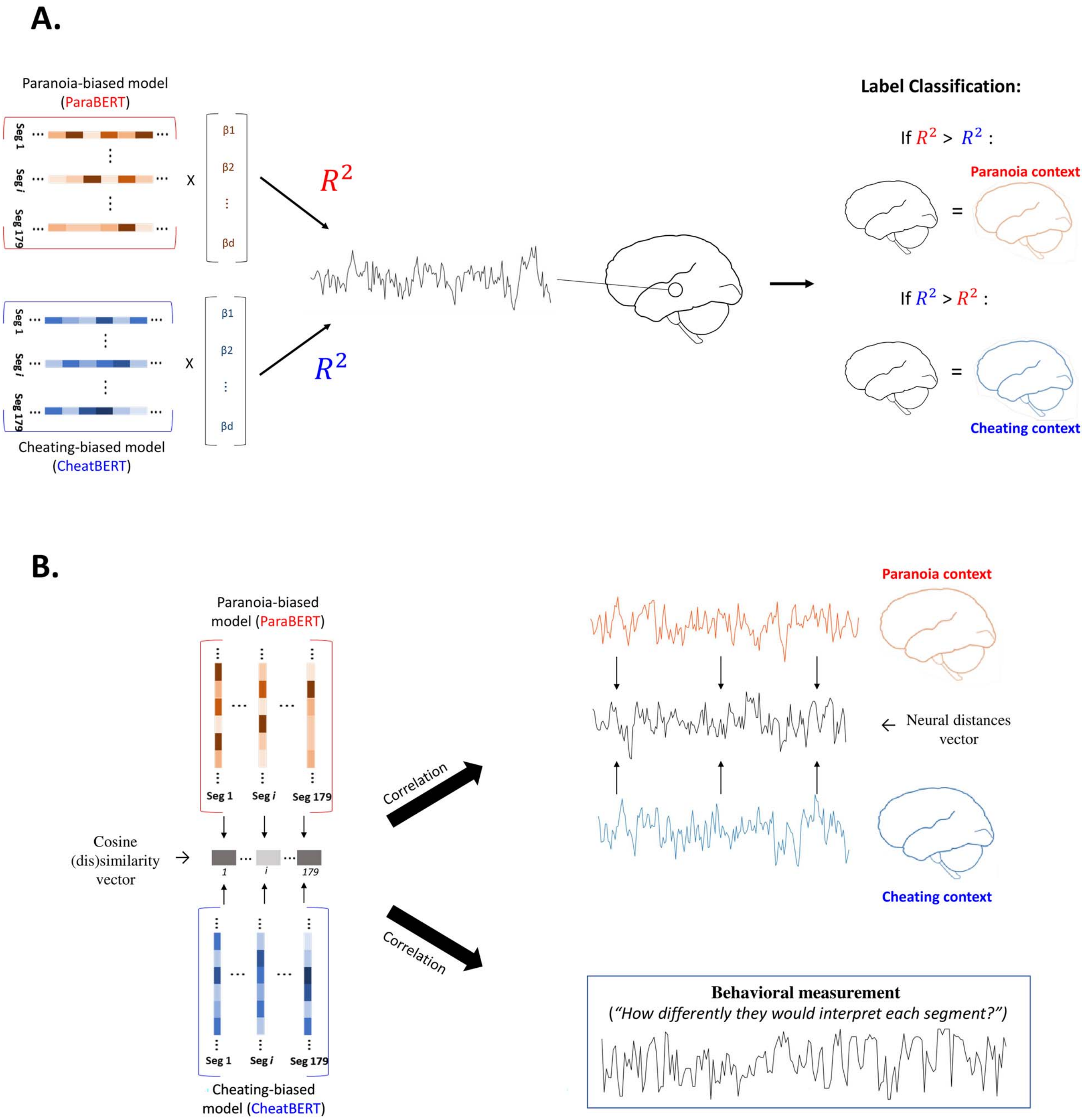

Fig. 4. Illustrations of the two primary analyses applied in this paper. A) The voxel-wise encoding-based classifier. The classifier predicts the context in which subjects interpreted the story, by competing the models against each other in their ability to encode the voxel's BOLD signal. The context would classify as cheating if the CheatBERT-based encoder is better than the ParaBERT-based encoder, and vice versa (measured by comparing the encoders' R_adj2 scores). B) The distance analysis. The difference between the cheating-induced and the paranoia-induced interpretations was quantified for each segment, in all modalities: Language models, brains, and behavior. From the language models, we extracted a distance vector by calculating the cosine dissimilarity between models' vector representations. From the neural data, we extracted a distance vector for each voxel by taking the absolute values of the differences between the averaged signal of the cheating group and the averaged signal of the paranoia group. The behavioral distances vector was collected by asking five independent raters to rate, for each segment, how differently subjects from different groups would interpret the segment. We analyzed the correlations between the models' distance vector and the neural distance vectors, and between the models' distance vector and the behavioral measurement.

Distances between the word embedding spaces are correlated with both neural and behavioral distances

The effect of the listener's perspective on the narrative interpretations is not fixed, as some moments in the story are more ambiguous and malleable to shift in context, while others are less open to multiple interpretations. To assess whether our fine-tuned models capture the dynamic fluctuations in interpretability across subjects who listened to the same story while having two opposing perspectives (contexts) we performed two analyses.

In the first analysis, we calculated, for every segment, the cosine dissimilarity score between the vector representations of both CheatBERT and ParaBERT, and correlated these scores with the differences in the neural activity between the groups (See Materials and methods and Fig. 4B). Our analysis

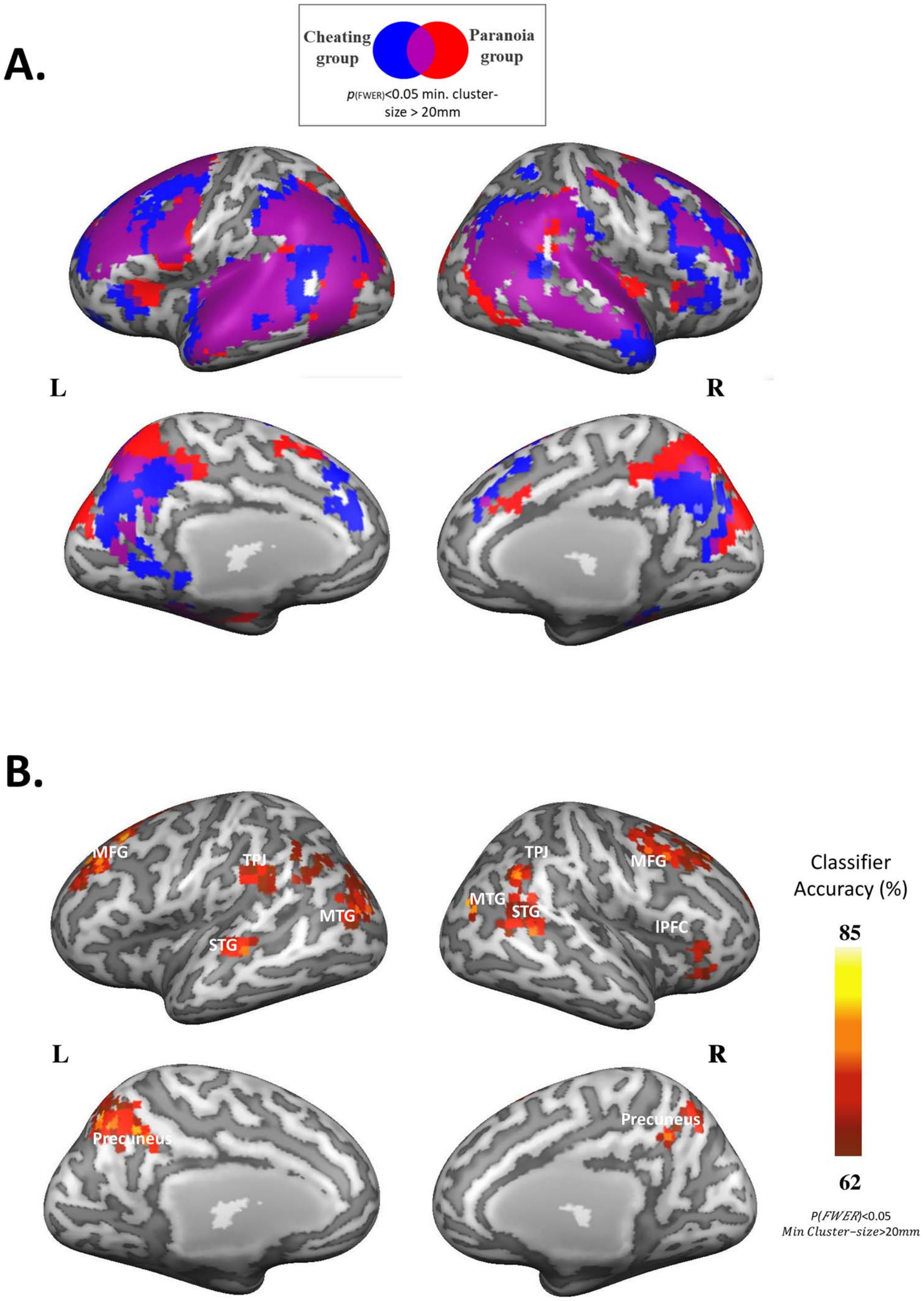

Fig. 5. A) Cortical maps showing voxels that are significantly encoded [i.e. with a significant R^2_{adj} score, p(FWER) < 0.05, minimum cluster-size >20 mm^2] by the fine-tuned models. Averaged R^2_{adj}s are 0.37 (min–max range: 0.26–0.59) for CheatBERT/cheating group, and 0.35 (min–max range: 0.24–0.59) for ParaBERT/paranoia group. For direct comparison between the fine-tuned models and the un-tuned BERT model, see Fig. S4. B) The accuracy scores map of the encoder-based classification analysis. The map contains only significant voxels [p(FWER) < 0.05, minimum cluster-size >20 mm^2].

revealed significant correlations (between $r = 0.3$ and $r = 0.43$, $p(\text{FWER}) < 0.05$) in extensive brain areas, including the bilateral TPJ, Precuneus, Premotor Cortex, STG, and Insula, as well as in the right MTG, right Anterior Cingulate Cortex, and the left Hippocampus (a total of 1,020 voxels, Fig. 6A and Table S2). Importantly, running the same analysis with other combinations of models (i.e. the pairs: BERT-CheatBERT, BERT-ParaBERT, MedBERT-SpaceBERT, GunsBERT-MideastBERT) did not reveal any significant correlation-maps (For BERT-ParaBERT, MedBERT-SpaceBERT and GunsBERT-MideastBERT we found zero significant voxels after correcting for multiple comparisons. For BERT—CheatBERT we did find 30 significant voxels, but they did not reach the cluster-size threshold.)

In the next analysis, we compared the difference in the representation of each segment by ParaBERT and CheatBERT (cosine dissimilarity vector, Fig. 4B) to the estimated change in interpretation between listeners exposed to cheating or the paranoia contexts (behavioral dissimilarity). Behavioral dissimilarity was assessed using independent raters that assessed ‘how different subjects in the cheating condition and in the paranoia condition

Table 1. Brain regions that showed significant accuracy scores in the voxel-wise encoder-based classifier (the first analysis).

				Coordinates		
Region	Hem.	No. of Voxels	Peak Accuracy score	X	Y	Z
Precuneus	Bilateral	243	0.85	−6	−60	39
Middle Frontal Gyrus	Right	120	0.725	35	−4	41
Middle Frontal Gyrus	Left	119	0.725	−31	15	43
Temporoparietal Junction	Right	69	0.725	39	−67	35
Temporoparietal Junction	Left	78	0.7	−42	−44	32
Middle Temporal Gyrus	Right	15	0.775	58	−47	10
Middle Temporal Gyrus	Left	67	0.7	−55	−52	3
Superior Temporal Sulcus	Right	78	0.7	53	−39	13
Superior Temporal Gyrus	Left	70	0.7	−53	−27	8
Ventrolateral Prefrontal Cortex	Right	62	0.65	37	19	11

would interpret each segment' (see Materials and methods and Fig. 4B). The CheatBERT-ParaBERT distance scores were significantly correlated with the behavioral scores ($r = 0.31$, $P < 0.001$). In contrast, the correlations between the distances between control models (i.e. the MedBERT-SpaceBERT distance and the GunsBERT-MideastBERT distance) and the behavioral scores were significantly lower ($r = 0.15$ and $r = 0.08$ for MedBERT-SpaceBERT and GunsBERT-MideastBERT, respectively) and not significantly different from zero ($P > 0.05$ for both, Fig. 6B).

Discussion

We presented a computational framework for modeling the effect of listeners' perspective, as defined by their thoughts, beliefs, and knowledge, on the way they interpret the exact same lingual stimulus. Recent research found similarities between word embedding representations derived from DLMs and language-related signals of the human brain (Huth et al. 2016; Jain and Huth 2018; Pereira et al. 2018; Gauthier and Levy 2019; Schwartz et al. 2019; Caucheteux et al. 2021; Goldstein et al. 2022a). Although word embedding representations are sensitive to textual context (the representation of a single word might be changed as a function of the other words in the sentence) they are not sensitive to extratextual information, such as the listener's thoughts, beliefs, and prior knowledge (i.e. the listener's perspective). To bridge this gap, we proposed a computational mechanism by which we create multiple word embedding spaces, each suitable for language comprehension under a different extratextual context. We suggested that among humans, the extratextual context (i.e. the listener's perspective) induces transformations in the neural representation of the language, and that the alternation between several carefully designed word embedding spaces can model neural changes in human listeners caused by different perspectives.

To create multiple word embedding spaces we collected a dedicated dataset (Cheating/Paranoia/Natural stories) and used it to fine-tune a well-established DLM, BERT (Devlin et al. 2018), to fit either the 'cheating' or the 'paranoia' extratextual contexts. We consistently showed that the fine-tuned DLMs (CheatBERT and ParaBERT) better fit the neural responses of the subjects with the corresponding perspective (Cheating vs. Paranoia). First, we showed that we can use fine-tuned models' word embeddings to successfully predict the context in which a subject interpreted the story (Figs. 4A and 5). Second, we found that the magnitude of change in the representation of each segment between ParaBERT and CheatBERT (measured via cosine dissimilarity) is correlated with the magnitude of change in neural responses between the cheating and the paranoia groups (Figs. 4B and 6A).

Besides the association with the neural data, we investigated the differences between the word embedding spaces of CheatBERT and ParaBERT and found that they are reasonably consistent with the human perspectives. The cosine dissimilarity score of each segment, as calculated between the fine-tuned models, was not only correlated with the magnitude of change in the neural responses (Figs. 4B and 6A) but also with the expected level of difference in the interpretation, as rated by independent raters (Fig. 6B). Moreover, we focused on the 20 most perspective-affected segments and showed that their word representations, especially the representations of a-priori-selected keywords, were significantly changed between CheatBERT and ParaBERT, compared to the other words in the story, as well as the other control models' representations (Fig. 3, Table S3). Taken all together, our results demonstrate how aligning a continuous embedding space to a specific context can provide a novel way of modeling listeners' intrinsic perspectives.

It is important to emphasize that the fine-tuning process itself does not necessarily reflect the way the brain switches between internal representations, instead it is only our technical way to induce the relevant embedding spaces. That is, we do not aim to describe the way the embedding space should be changed, but instead we argue that when the embedding spaces of DLMs are properly tuned, they capture brain activity information about the extratextual context, and this is an evidence of a potential cognitive mechanism of handling such a context. Future work is needed in order to model not just the differences between representations (word embedding spaces), as we successfully demonstrated here, but also the computational transformation that causes these changes.

It should be also noted that when analyzing the differences between the fine-tuned models using cosine dissimilarity (Figs. 2, 4B, and 6), we must assume that the word embedding vectors of the two models are directly comparable, despite the fact that fine-tuning may arbitrarily alter the original meaning of the embedding-space dimensions. The above results suggest that this assumption holds in our study, as we show that these cosine-dissimilarity scores are predictive of both the behavioral and the neural data. Moreover, a careful analysis of the word embedding representations of the Salinger's story reveals that the vector representations of the two fine-tuned models did not deviate much from the representations of the original BERT model, in terms of cosine similarity (The cosine-similarity scores between CheatBERT's and BERT's representations were between 0.84

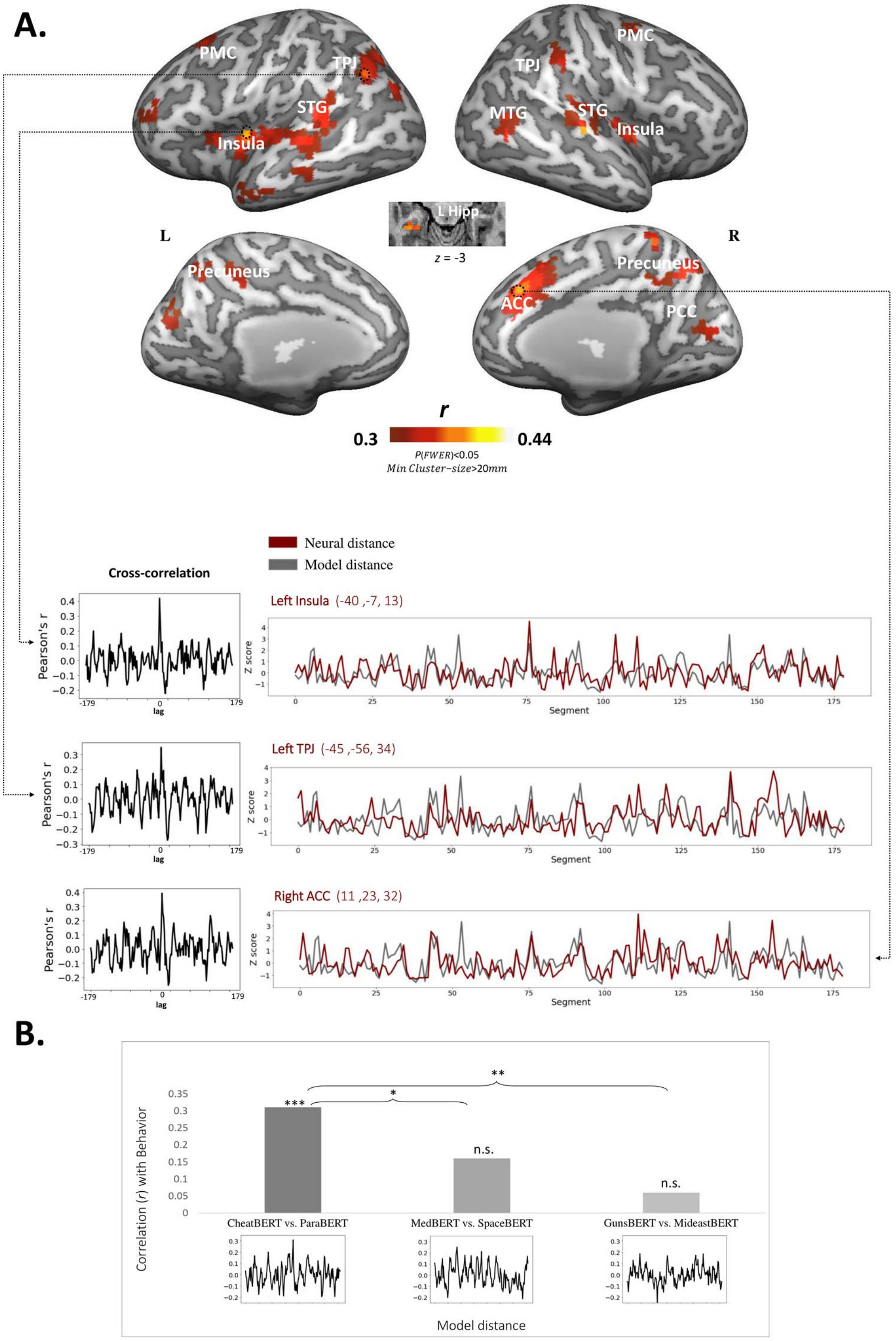

Fig. 6. Results of the distance analysis. A) A correlation map showing voxels whose neural distances were significantly correlated with model distance [p(FWER) < 0.05, minimum cluster-size >20mm^2]. For several brain regions we plot the neural distance fluctuations as measured across segments (maroon colored line), together with the models' distances (gray colored line). A cross-correlation plot is attached to the plot of each of the regions to visually indicate the signal-to-noise ratio Hipp = hippocampus. B) A bar-plot showing the correlations (Pearson's r) between model distances—As extracted from different pairs of models—And the behavioral scores. Below each bar is the corresponding cross-correlation plot. *P < 0.05, **P < 0.01, ***P < 0.005, n.s. = non-significant (P > 0.05).

and 0.97; The cosine-similarity scores between ParaBERT's and BERT's representations were between 0.78 and 0.94). This implies that our fine-tuning process effectively preserved most of the linguistic information and made only minimal changes to the representations- changes that successfully captured the variations in human perspectives. Lastly, even if we don't assume that the two models share the same space, we still consider direct comparison between them to be reasonable, as we can attribute the distance between the models' representations to the extent each model's representation has changed with respect to the original BERT. In other words, when the representation of a word only undergoes minimal changes from the original

BERT representation in both fine-tuned models (such as a neutral word not related to cheating or paranoia), the distance between the fine-tuned models for that word should also be minimal. Conversely, the more relevant a word is to the cheating/paranoia context, the more it deviates from the BERT representation, leading to a greater likelihood of the new fine-tuned models' representations being far apart from each other. Indeed, in our study, the distances between the representations of the two fine-tuned models are empirically correlated with the distances between each fine-tuned model's representations and the original BERT's representations. The correlation between the distance scores of CheatBERT vs. ParaBERT and the distance scores of CheatBERT vs. BERT was 0.58 ($P < 0.001$). Similarly, the correlation between the distance scores of CheatBERT vs. ParaBERT and the distance scores of ParaBERT vs. BERT was 0.69 ($P < 0.001$).

In this study, we implemented our fine-tuning procedure using the BERT model, a classical model that was originally designed to be trained via a two-stage procedure: pre-training followed by fine-tuning. Although recent studies suggest that new models, such as GPT-2 (Radford et al. 2019), that are based on a decoder architecture, are more predictive of neural responses to language stimuli than encoder-based models like BERT (Caucheteux et al. 2021; Schrimpf et al. 2021), the current methodological framework of using fine-tuning is less compatible with decoder-based models (Radford et al. 2019, Liu et al. 2021; Stylianou and Vlahavas 2021; Winata et al. 2021; Wortsman et al. 2022). Indeed, our supplementary analysis (Fig. S5) reveals that our results are successfully replicated using a different encoder-based model (RoBERTa; Liu et al. 2019) while they are replicated less successfully when applying the fine-tuning procedure to the GPT-2 model.

From a computational point of view, the current study gives a new perspective on the concept of fine-tuning. Usually, fine-tuning is intended for improving the performance of the DLM in downstream tasks, as text summarization, machine translation, and so on. In other words, the fine-tuning stage creates a unique and dedicated variant of the DLM for each downstream task (or 'stimulus'). Here, however, we used a different logic: instead of adopting a single model to a stimulus, we use fine-tuning to create multiple models that will later be applied to model the same stimulus (i.e. Salinger's story) and examine differences in how listeners with different perspectives perceive them. This approach has a great advantage in cognitively motivated computational modeling, since in real life, we may process the same stimulus in different ways, depending on our given state-of-mind.

Supplementary material

Supplementary material is available at *Cerebral Cortex* online.

Conflict of interest statement: The authors declare no competing financial interests.

References

Adolphs R. The social brain: neural basis of social knowledge. *Annu Rev Psychol*. 2009:60:693–716.

Aharoni R, Johnson M , Firat O. Massively multilingual neural machine translation. 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), Minneapolis, Minnesota. Association for Computational Linguistics; 2019. p. 3874–3884.

Caucheteux C, Gramfort A, King JR. Model-based analysis of brain activity reveals the hierarchy of language in 305 subjects. In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic. Association for Computational Linguistics; 2021. p. 3635–3644.

Devlin J, Chang MW, Lee K, Toutanova K. 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), Minneapolis, Minnesota. Association for Computational Linguistics; 2019. p. 4171–4186.

Federmeier KD, Segal JB, Lombrozo T, Kutas M. Brain responses to nouns, verbs and class-ambiguous words in context. *Brain*. 2000:123(12):2552–2566.

Fletcher PC, Happe F, Frith U, Baker SC, Dolan RJ, Frackowiak RS, Frith CD. Other minds in the brain: a functional imaging study of "theory of mind" in story comprehension. *Cognition*. 1995:57(2): 109–128.

Gauthier J, Levy R. Linking artificial and human neural representations of language. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China. Association for Computational Linguistics; 2019. p. 529–539.

Goldstein A, Dabush A, Aubrey B, Schain M, Nastase SA, Zada Z, Hasson U. Brain embeddings with shared geometry to artificial contextual embeddings, as a code for representing language in the human brain. BioRxiv pre-print. 2022b:2022–2003. Retrieved from https://www.biorxiv.org/content/10.1101/2022.03.01.482586v1.

Goldstein, A., Zada, Z., Buchnik, E., et al. Shared computational principles for language processing in humans and deep language models. *Nat Neurosci* 2022a:25:369–380. https://doi.org/10.1038/s41593-022-01026-4.

Hasson U, Nir Y, Levy I, Fuhrmann G, Malach R. Intersubject synchronization of cortical activity during natural vision. *Science*. 2004:303(5664):1634–1640.

Hinton GE. Learning multiple layers of representation. *Trends Cogn Sci*. 2007:11(10):428–434.

Huth AG, De Heer WA, Griffiths TL, Theunissen FE, Gallant JL. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*. 2016:532(7600):453–458.

Jain S, Huth AG. Incorporating context into language encoding models for fMRI. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. Curran Associates, Inc., Montreal, Canada; 2018. p. 6629–6638.

Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Stoyanov V. Roberta: a robustly optimized bert pretraining approach. arXiv preprint arXiv:190711692. 2019. Retrieved from https://arxiv.org/abs/1907.11692.

Liu X, Zheng Y, Du Z, Ding M, Qian Y, Yang Z, Tang J. GPT understands, too. arXiv preprint arXiv:210310385. 2021. Retrieved from https://arxiv.org/abs/2103.10385.

Mar RA. The neural bases of social cognition and story comprehension. *Annu Rev Psychol*. 2011:62:103–134.

Mars RB, Neubert FX, Noonan MP, Sallet J, Toni I, Rushworth MF. On the relationship between the "default mode network" and the "social brain". *Front Hum Neurosci*. 2012:6:189.

Nichols T, Hayasaka S. Controlling the familywise error rate in functional neuroimaging: a comparative review. *Stat Methods Med Res*. 2003:12(5):419–446.

Nieuwland MS. "Who's he?" event-related brain potentials and unbound pronouns. *J Mem Lang*. 2014:76:1–28.

Pappagari R, Zelasko P, Villalba J, Carmiel Y, Dehak N. Hierarchical transformers for long document classification. In: *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. Singapore; 2019. pp. 838–844. https://doi.org/10.1109/ASRU46091.2019.9003958.

Pereira F, Lou B, Pritchett B, Ritter S, Gershman SJ, Kanwisher N, Fedorenko E. Toward a universal decoder of linguistic meaning from brain activation. *Nat Commun*. 2018:9(1):1–13.

Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L. Deep contextualized word representations. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long Papers), New Orleans, Louisiana. Association for Computational Linguistics; 2018. p. 2227–2237.

Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. *OpenAI blog*. 2019:1(8):9.

Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*. 2020:21(1):5485–5551.

Rogers A, Kovaleva O, Rumshisky A. A primer in bertology: what we know about how bert works. *Trans Assoc Comput Linguistics*. 2020:8: 842–866.

Schrimpf M, Blank IA, Tuckute G, Kauf C, Hosseini EA, Kanwisher N, Tenenbaum JB, Fedorenko E. The neural architecture of language: Integrative modeling converges on predictive processing. In: *Proceedings of the National Academy of Sciences*. 2021:118(45). https://doi.org/10.1073/pnas.2105646118.

Schwartz D, Toneva M, Wehbe L. Inducing brain-relevant bias in natural language processing models. *Adv Neural Inf Proces Syst*. 2019:32:14123–14133 .

Simony E, Honey CJ, Chen J, Lositsky O, Yeshurun Y, Wiesel A, Hasson U. Dynamic reconfiguration of the default mode network during narrative comprehension. *Nat Commun*. 2016:7(1):1–13.

Stylianou N, Vlahavas I. CoreLM: Coreference-aware language model fine-tuning. In: *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*. Punta Cana, Dominican Republic; 2021. p. 70–81.

Talairach J, Tournoux P. *Co-planar stereotaxic atlas of the human brain* (M. Rayport, Trans.). New York, NY: Thieme Medical Publishers; 1988.

Van Berkum JJ. Understanding sentences in context: what brain waves can tell us. *Curr Dir Psychol Sci*. 2008:17(6):376–380.

Van Berkum JJ, Nieuwland MS. A cognitive neuroscience perspective on language comprehension in context. In: *Human Language: From Genes and Brain to Behavior*. MIT Press; 2019. pp. 429–442.

Winata GI, Madotto A, Lin Z, Liu R, Yosinski J, Fung P. Language models are few-shot multilingual learners. In: *Proceedings of the 1st Workshop on Multilingual Representation Learning*, Punta Cana, Dominican Republic: Association for Computational Linguistics; 2021. p. 1–15.

Wortsman M, Ilharco G, Kim JW, Li M, Kornblith S, Roelofs R, Lopes RG, Hajishirzi H, Farhadi A, Namkoong H, et al. Robust fine-tuning of zero-shot models. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, LA, USA, 2022, pp. 7949–7961. https://doi.org/10.1109/CVPR52688.2022.00780.

Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, Dean J. Google's neural machine translation system: bridging the gap between human and machine translation. arXiv preprint arXiv:160908144. 2016. Retrieved from https://arxiv.org/abs/1609.08144.

Yeshurun Y, Swanson S, Simony E, Chen J, Lazaridi C, Honey CJ, Hasson U. Same story, different story: the neural representation of interpretive frameworks. *Psychol Sci*. 2017:28(3): 307–319.

Yeshurun Y, Nguyen M, Hasson U. The default mode network: where the idiosyncratic self meets the shared social world. *Nat Rev Neurosci*. 2021:22(3):181–192.