Generating structured stimuli for investigations of human behavior and brain activity with computational models

Matthew E. Siegelman

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2024

Abstract

Generating structured stimuli for investigations of human behavior and brain activity with computational models

Matthew E. Siegelman

Some of the most important discoveries in cognitive neuroscience have come from recent innovations in experimental tools. Computational models that simulate human perception of environmental inputs have revealed the internal processes and features by which those inputs are learned and represented by the brain. We advance this line of work across two separate research studies in which we leveraged these models to both generate experimental task stimuli and make predictions about behavioral and neural responses to those stimuli. Chapter 1 details how nine language models were used to generate *controversial sentence pairs* for which two of the models disagreed about which sentence is more likely to occur. Human judgments about these sentence pairs were collected and compared to model preferences in order to identify model-specific pitfalls and provide a behavioral performance benchmark for future research. We found that transformer models GPT-2, RoBERTa and ELECTRA were most aligned with human judgments. Chapter 2 utilizes the GloVe model of semantic word vectors to generate a set of schematically structured poems comprising ten different topics whose specific temporal order was learned by a group of participants. The GloVe model was then used to investigate learning-induced changes in the spatial geometry of the representations of the topics across the cortex. A Hidden Markov Model was also used to measure neural event segmentation during poem listening. In both analyses we identified a consistent topography of learning-induced changes in the default mode network, which could be partially explained by the models.

Table of Contents

List of Tables and Figures

Acknowledgments

To Mom: Thank you for everything. Looking forward to Europe this Fall.

To Dad: A brilliant docta, and a better fatha.

To Dan and Dyl: To quote Dan's camp counselor – "Brothers, I love you."

To Chris: I could fill this page listing your incredible qualities as a mentor. You are extremely knowledgeable, generous, and patient. I learned more from you personally than from all my graduate courses combined. I have really enjoyed being excited about ideas with you. I have no doubt you will continue to produce seminal work and train many others to do the same. The field is lucky to have you. I was lucky to have you.

To Mariam: Thank you for embracing the Alyssano logo. I still feel proud of that.

To Niko: Your talk at MIT in 2017 is the reason I took interest in Columbia.

To Tal: Thank you for introducing me to controversial stimuli, and for driving our project home.

To Ev: You taught me to turn words into numbers. Your influence runs deep in this dissertation.

To my friends: Thank you for pretending to want to read this.

To Bel: Thinking about you right now.

Introduction

Why do psychologists want to construct a model the brain? To better understand how it works. Our best neuroimaging methods have a limited spatial and temporal resolution. A model that appears to match the behavior and activity of the brain could help us gain deeper insights into the kinds of computations the brain uses to process information. How do we test a computational model of the brain? The same way we test the brain – by recording its output and internal activity in response to a certain stimulus. If the model acts like the brain, it could simulate the neural computations and representations of that stimulus.

For example, we can show pictures to deep convolutional neural networks (CNNs). These models have several hierarchical processing layers that integrate information over many spatial scales, which allows them to recognize objects in images (Kriegeskorte, 2015). The output of these models is a probabilistic label for the objects in the images, and the internal activity is the activation of units within each processing layer of the model. The human visual cortex also has several hierarchical processing layers, which also integrate information over different spatial scales to allow us to recognize objects. And in fact, it has been shown that the kinds of visual features relevant to human behavior and represented in the visual cortex are similar to those in deep CNNs (Kriegeskorte, 2015; Krizhevsky et al., 2017).

Just as CNNs integrate information over different spatial scales, recurrent neural network models (RNNs) integrate information over different time scales, allowing them to recognize patterns in language inputs (Bengio et al., 2000). The output of these models is a probabilistic prediction of an upcoming word, and the internal activity is embedded in its recurrent internal state, which updates with each new word. Also like CNNs, RNNs have helped us learn about the brain. By comparing the internal activity of RNNs to neural activity in the language system

during language comprehension, psychologists have found regions in the temporal lobe that integrate linguistic information over short and long timescales (Jain et al., 2020). More recent transformer models, such as BERT and GPT (Devlin et al., 2019; Radford et al., 2019), have been used to predict fine-grained brain responses in the language system during sentence comprehension (Schrimpf et al., 2021).

In all such cases of using models to understand the brain, the original source of information was human ecology, where natural human behavior created training data for these models. We took pictures of our experiences and uploaded them to the internet. We communicated with each other online with digital text. Models were then architected specifically to capture the structure of these multiple modalities of experience, and in turn helped us learn about our own brains and behavior. This is a compelling story, not only for the computational model architects and cognitive neuroscientists who have helped advance this knowledge, but for the billions of people on this planet who unwittingly contributed rich data sources to this scientific enterprise, and who benefit directly from its labors. Today, models like Dall-E and chatGPT are widely used across an array of business, artistic and personal ventures.

All that said, using these models to better understand the brain does have some limitations. For one thing, machine learning models tend to have strange, non-humanlike properties. The best object recognition models are routinely fooled by adversarial examples (Goodfellow et al., 2015), where random noise added to the pixel values of an image can cause a model to become extremely confident that an incorrect, unrelated object is present in an image. RNNs and transformer models have widely identified pitfalls, such as endless word and sentence-level repetitions characterized by a self-reinforcement effect (Xu et al., 2022). These models have also been shown to be capable of fooling each other into thinking that an obviously

errored sentence is relatively more likely than a human-preferred sentence (Golan et al., 2023). These odd behaviors suggest that there are still critical misalignments between models and human brains.

Another limitation to this approach is that good models tend to act similarly in response to typical inputs. Our field has been saturated with so many well performing models, in part because anytime a research lab achieves 0.1% greater accuracy than the current leader on a popular benchmark task, it is deemed worthy of publication. How do we compare all these models to each other, or determine which model is acting similarly to human behavior or the brain, when they all converge on similar solutions and respond in a similar manner? We need new methods to make models behave differently in interesting ways.

Finally, even the most sophisticated models do not fully capture naturalistic human experiences. Models have proven effective at simulating responses to things like basic linguistic and visual tasks, which have bottom up, ground truth structure. When we aim to study higher level brain functions and behaviors that involve dynamic perception, memory, prediction, emotion, and attention, we must use stimuli that match the richness and complexity of the real word, such as movies and stories. However, these types of complex experiences are not captured by our current models.

In this dissertation, we propose that one way to push against these various limitations is to *use these models to generate stimuli for experiments*. By doing so, we can create specially crafted inputs that attempt to "break" the alignment between models and typical human behavior, so that if a model does have a so-called "strange property," it can be identified in an efficient way. We can also identify stimuli that optimally distinguish between models which tend to behave similarly and design stimulus sets to exploit those distinctions, thus introducing

meaningful variance between the models to then compare them to humans. Lastly, we can guide model outputs to create structured stimuli of intermediate complexity, which are engaging for participants by incorporating naturalistic properties such as temporal dynamics and semantic complexity, and yet are also tied directly to model features and can therefore be used to test-model driven hypotheses about higher level cognition.

The work reported in this dissertation utilizes that line of reasoning. In both chapters, we use language models to generate highly structured experimental task stimuli, and then input the stimuli back into the models to make specific predictions about new human data. In Chapter 1, we used nine different language models (including n-gram, recurrent neural networks, and transformers) to construct *controversial sentence pairs* for which two of the models disagreed about which sentence is more likely to occur. We compared model preferences to human judgments on the same set of sentence pairs, including sentences sampled from natural text, and synthetic sentences optimized to be controversial for a given pair of models. We found that 1) GPT-2 (a unidirectional transformer model trained on predicting upcoming tokens) and RoBERTa (a bidirectional transformer trained on a held-out token prediction task) were the most predictive of human judgments on controversial natural sentence pairs; 2) GPT-2, RoBERTa, and ELECTRA (a bidirectional transformer trained on detecting corrupted tokens) were the most predictive of human judgments on synthetic sentence pairs; and 3) GPT-2 was the most human-consistent model when considering the entire behavioral dataset we collected. These findings coincide with recent evidence that transformers also outperform recurrent networks for predicting behavioral reading speed (Wilcox et al., 2021; Merkx & Frank, 2021) and neural responses to natural language (Schrimpf et al., 2021; Goldstein et al., 2022). However, all of the models, including GPT-2, exhibited behavior inconsistent with human judgments; using an

alternative model as a counterforce, we could corrupt natural sentences such that their probability under a model did not decrease, but humans tended to reject the corrupted sentence as unlikely. These errors could be informative about model-specific pitfalls related to their architectures or training procedures. They also provide a behavioral performance benchmark to compare to future neurobiologically plausible models, which might be more robust to this type of corruption. Those models could then be probed further, either by pitting them against each other as we showed here, or by synthesizing stimuli for another experiment.

In Chapter 2, we used BERT (a bidirectional transformer) and GloVe (a non-recurrent model of semantic features) to generate a set of schematically structured poems comprising a temporal sequence of ten different *topics*. We used these stimuli to investigate changes in the temporal dynamics and semantic representations associated with learning a new *schema* by playing spoken recordings of the poems in fMRI before and after participants learned its topic sequence. We found a topography of changes that was consistent between two independent groups of participants who learned two different sequences across multiple regions of the default mode network (DMN), in line with previous work on schema perception and memory (Baldassano et al., 2017; Baldassano et al., 2018). In a Hidden Markov Model analysis, we showed how changes in the neural time course activity associated with individual poems could be explained in one group of participants by an increase in the strength of *event boundaries* – shifts between stable patterns of neural activity at event transitions – which facilitate the mental segmentation of experiences in perception and memory (Kurby & Zacks, 2008). In a representational similarity analysis, we used the GloVe model, which has been previously used to decode semantic representations in fMRI data in response to a wide variety of concrete and abstract topics (Pereira et al., 2018), to describe how topic representations in the DMN changed

with respect to their spatial geometry in GloVe vector space, thereby making full use of this model by both constructing experimental task stimuli and subsequently predicting brain responses to those stimuli.

The results reported in these studies provide novel insights into current debates in the literature and create unique opportunities for further investigation. In both chapters we build on recent work applying similar experimental procedures and introduce several novel methods that could materially benefit future research.

Chapter 1: Testing the limits of natural language models for predicting human language judgments

Please note, chapter to be published as:

Golan, T.*, **Siegelman, M.***, Kriegeskorte, N. & Baldassano, C. (Accepted). Testing the limits of natural language models for predicting human language judgments. *Nature Machine Intelligence.*

∗The first two authors contributed equally to this work.

1.1 Introduction

Neural network language models are not only key tools in natural language processing (NLP) but are also drawing an increasing scientific interest as potential models of human language processing. Ranging from recurrent neural networks (Rumelhart et al., 1986; Hochreiter & Schmidhuber, 1997) to transformers (Devlin et al., 2019; Liu et al., 2019; Conneau & Lample, 2019; Clark et al., 2020; Radford et al., 2019), each of these language models (explicitly or implicitly) defines a probability distribution over strings of words, predicting which sequences are likely to occur in natural language. There is substantial evidence from measures such as reading times (Goodkind & Bicknell, 2018), functional MRI (Shain et al., 2020), scalp EEG (Broderick et al., 2018), and intracranial ECoG (Goldstein et al., 2022) that humans are sensitive to the relative probabilities of words and sentences as captured by language models, even among sentences that are grammatically correct and semantically meaningful. Furthermore, model-derived sentence probabilities can also predict human graded acceptability judgments (Lau et al., 2017; Lau et al., 2020). These successes, however, have not yet addressed two central questions of interest: (1) Which of the models is best-aligned with human language processing? (2) How close is the best-aligned model to the goal of fully capturing human judgments?

The standard approach for evaluating language models is to use a set of standardized benchmarks such as those in the General Language Understanding Evaluation (GLUE) (Wang et al., 2019a), or its successor, SuperGLUE (Wang et al., 2019b). Though instrumental in evaluating the utility of language models for downstream NLP tasks, these benchmarks prove insufficient for comparing such models as candidate explanations of human language-processing. Many components of these benchmarks do not aim to measure human alignment, but rather the usefulness of the models' language representation when tuned to a specific downstream task.

Some benchmarks challenge language models more directly by comparing the probabilities they assign to grammatical and ungrammatical sentences (e.g., Warstadt et al., 2020). However, since such benchmarks are driven by theoretical linguistic considerations, they might fail to detect novel, unexpected ways in which language models may diverge from human language understanding. Last, an additional practical concern is that the rapid pace of NLP research has led to rapid saturation of these kinds of static benchmarks, making it difficult to distinguish between models (Kiela et al., 2021).

One proposed solution to these issues is the use of dynamic human-in-the-loop benchmarks in which people actively stress-test models with an evolving set of tests. However, this approach faces the major obstacle that "finding interesting examples is rapidly becoming a less trivial task" (Kiela et al., 2021). We propose to complement human-curated benchmarks with model-driven evaluation. Guided by model predictions rather than experimenter intuitions, we would like to identify particularly informative test sentences, where different models make divergent predictions. This approach of running experiments mathematically optimized to "put in jeopardy" particular models belongs to a long-standing scientific philosophy of design optimization (Box & Hill, 1967). We can find these critical sentences in large corpora of natural language or synthesize novel test sentences that reveal how different models generalize beyond their training distributions.

We propose here a systematic, model-driven approach for comparing language models in terms of their consistency with human judgments. We generate controversial sentence pairs: pairs of sentences designed such that two language models strongly disagree about which sentence is more likely to occur. In each of these sentence pairs, one model assigns a higher probability to the first sentence than the second sentence, while the other model prefers the

second sentence to the first. We then collect human judgments of which sentence in each pair is more probable to settle this dispute between the two models.

This approach builds on previous work on controversial images for models of visual classification (Golan et al., 2020). That work relied on absolute judgments of a single stimulus, which are appropriate for classification responses. However, asking the participants to rate each sentence's probability on an absolute scale is complicated by between-trial context effects common in magnitude estimation tasks (Cross, 1973; Foley et al., 1990; Petzschner et al., 2015), which have been shown to impact judgments like acceptability (Greenbaum, 1977). A binary forced-choice behavioral task presenting the participants with a choice between two sentences in each trial, the approach we used here, minimizes the role of between-trial context effects by setting an explicit local context within each trial. Such an approach has been previously used for measuring sentence acceptability (Schutze & Sprouse, 2014) and provides substantially more statistical power compared to designs in which acceptability ratings are provided for single sentences (Sprouse & Almeida, 2017).

Our experiments demonstrate that 1) it is possible to procedurally generate controversial sentence pairs for all common classes of language models, either by selecting pairs of sentences from a corpus or by iteratively modifying natural sentences to yield controversial predictions; 2) the resulting controversial sentence pairs enable efficient model comparison between models that otherwise are seemingly equivalent in their human consistency; and 3) all current NLP model classes incorrectly assign high probability to some non-natural sentences (one can modify a natural sentence such that its model probability does not decrease but human observers reject the sentence as unnatural). This framework for model comparison and model testing can give us new

insight into the classes of models that best align with human language perception and suggest directions for future model development.

1.2 Methods

Language models

We tested nine models from three distinct classes: n-gram models, recurrent neural networks, and transformers. The n-gram models were trained with open source code from the Natural Language Toolkit (Bird et al., 2009), the recurrent neural networks were trained with architectures and optimization procedures available in PyTorch (Paszke et al., 2019), and the transformers were implemented with the open-source repository HuggingFace (Wolf et al., 2020). For full details see Supplementary Methods.

Evaluating sentence-level probabilities in transformer models

We then sought to compute the probability of arbitrary sentences under each of the models described above. The term "sentence" is used in this context in its broadest sense–a sequence of English words, not necessarily restricted to grammatical English sentences. Unlike some classification tasks in which valid model predictions may be expected only for grammatical sentences (e.g., sentiment analysis), the sentence probability comparison task is defined over the entire domain of eight-word sequences.

For the set of unidirectional models, evaluating sentence probabilities was performed simply by summing the log probabilities of each successive token in the sentence from left to right, given all the previous tokens. For bidirectional models, this process was not as straightforward. One challenge is that transformer model probabilities do not necessarily reflect a coherent joint probability; the summed log sentence probability resulting from adding words in one order (e.g. left to right) does not necessarily equal the probability resulting from a different

order (e.g. right to left). Here we developed a novel formulation of bidirectional sentence probabilities in which we considered all permutations of serial word positions as possible construction orders (analogous to the random word visitation order used to sample serial reproduction chains, Yamakoshi et al., 2022). In practice, we observed that the distribution of log probabilities resulting from different permutations tends to center tightly around a mean value (for example, for RoBERTa evaluated with natural sentences, the average coefficient of variation was approximately 0.059). Therefore in order to efficiently calculate bidirectional sentence probability, we evaluate 100 different random permutations and define the overall sentence log probability as the mean log probability calculated from each permutation. Specifically, we initialized an eight-word sentence with all tokens replaced with the "mask" token used in place of to-be-predicted words during model training. We selected a random permutation P of positions 1 through 8, and started by computing the probability of the word at first of these positions P_1 given the other seven "mask" tokens. We then replaced the "mask" at position P_1 with the actual word at this position and computed the probability of the word at P_2 given the other six "mask" tokens and the word at P_1. This process was repeated until all "mask" tokens had been filled by the corresponding word.

A secondary challenge in evaluating sentence probabilities in bidirectional transformer models stems from the fact that these models use word-piece tokenizers (as opposed to whole words), and that these tokenizers are different for different models. For example, one tokenizer might include the word "beehive" as a single token, while others strive for a smaller library of unique tokens by evaluating "beehive" as the two tokens "bee" and "hive". The model probability of a multi-token word–similar to the probability of a multi-word sentence–may depend on the order in which the chain rule is applied. Therefore, all unique permutations of

token order for each multi-token word were also evaluated within their respective "masks". For example, the probability of the word "beehive" would be evaluated as follows (Eq. 1):

$$\log p(w = \text{beehive}) = 0.5 \log p(w_1 = \text{bee} \mid w_2 = \text{MASK}) + \log p(w_2 = \text{hive} \mid w_2 = \text{bee}) + 0.5 \log p(w_2 = \text{hive} \mid w_1 = \text{MASK}) + \log p(w_1 = \text{bee} \mid w_2 = \text{hive})$$

This procedure aimed to yield a more fair estimate of the conditional probabilities of word-piece tokens and therefore the overall probabilities of multi-token words by 1) ensuring that the word-piece tokens were evaluated within the same context of surrounding words and masks, and 2) eliminating the bias of evaluating the word-piece tokens in any one particular order in models which were trained to predict bidirectionally.

One more procedure was applied in order to ensure that all models were computing a probability distribution over sentences with exactly 8 words. When evaluating the conditional probability of a masked word in models with word-piece tokenizers, we normalized the model probabilities to ensure that only single words were being considered, rather than splitting the masked tokens into multiple words. At each evaluation step, each token was restricted to come from one of four normalized distributions: i) single-mask words were restricted to be tokens with appended white space, ii) masks at the beginning of a word were restricted to be tokens with preceding white space (in models with preceding white space, e.g. BERT), iii) masks at the end of words were restricted to be tokens with trailing white space (in models with trailing white space, e.g. XLM), and iv) masks in the middle of words were restricted to tokens with no appended white space.

Assessing potential token count effects on sentence probability estimates

Note that, because tokenization schemes varied across models, the number of tokens in a sentence could differ for different models. These alternative tokenizations can be conceived of as different factorizations of the modeled language distribution, changing how a sentence's log

probability is additively partitioned across the conditional probability chain (but not affecting its overall probability) (Chestnut, 2019). Had we attempted to normalize across models by dividing the log probability by the number of tokens, as is done when aligning model predictions to human acceptability ratings (Lau et al., 2017; Lau et al., 2020), our probabilities would have become strongly tokenization-dependent (Chestnut, 2019). To empirically confirm that tokenization differences were not driving our results, we statistically compared the token counts of each model's preferred synthetic sentences with the token counts of their non-preferred counterparts. While we found significant differences for some of the models, there was no systematic association between token count and model sentence preferences (Table S2). In particular, lower sentence probabilities were not systematically confounded by higher token counts.

Defining a shared vocabulary

To facilitate the sampling, selection, and synthesis of sentences that could be evaluated by all of the candidate models, we defined a shared vocabulary of 29,157 unique words. Defining this vocabulary was necessary in order to unify the space of possible sentences between the transformer models (which can evaluate any input due to their word-piece tokenizers) and the neural network and n-gram models (which include whole words as tokens), and to ensure we only included words that were sufficiently prevalent in the training corpora for all models. The vocabulary consisted of the words in the subtlex database (Heuven et al., 2014), after removing words that occurred fewer than 300 times in the 300M word corpus used to train the n-gram and recurrent neural network models (i.e., with frequencies lower than one in a million).

Sampling of natural sentences

Natural sentences were sampled from the same four text sources used to construct the training corpus for the n-gram and recurrent neural network models (see above), while ensuring that there was no overlap between training and testing sentences. Sentences were filtered to include only those with eight distinct words and no punctuation aside from periods, exclamation points, or question marks at the end of a sentence. Then, all eight-word sentences were further filtered to include only the words included in the shared vocabulary and to exclude those included in a predetermined list of inappropriate words and phrases. To identify controversial pairs of natural sentences, we used integer linear programming to search for sentences that had above-median probability in one model and minimum probability rank in another model (see Supplementary Methods).

Generating synthetic controversial sentence pairs

For each pair of models, we synthesized 100 sentence triplets. Each triplet was initialized with a natural sentence n (sampled from Reddit). The words in sentence n were iteratively modified to generate a synthetic sentence with reduced probability according to the first model but not according to the second model. This process was repeated to generate another synthetic sentence from n, in which the roles of the two models were reversed. Conceptually, this approach resembles Maximum Differentiation (MAD) competition (Wang & Simoncelli, 2008), introduced to compare models of image quality assessment. Each synthetic sentence was generated as a solution for a constrained minimization problem (Eq. 2):

$$s^* = \underset{s}{\operatorname{argmin}} \log p(s \mid m_{reject})$$
$$\text{subject to } \log p(s \mid m_{accept}) \geq \log p(n \mid m_{accept})$$

m_{reject} denotes the model targeted to assign reduced sentence probability to the synthetic sentence compared to the natural sentence, and m_{accept} denotes the model targeted to maintain a

synthetic sentence probability greater or equal to that of the natural sentence. For one synthetic sentence, one model served as m_{accept} and the other model served as m_{reject}, and for the other synthetic sentence the model roles were flipped.

At each optimization iteration, we selected one of the eight words pseudorandomly (so that all eight positions would be sampled N times before any position would be sampled $N + 1$ times) and searched the shared vocabulary for the replacement word that would minimize the $\log p(s \mid m_{reject})$ under the constraint. We excluded potential replacement words that already appeared in the sentence, except for a list of 42 determiners and prepositions such as "the", "a", or "with", which were allowed to repeat. The sentence optimization procedure was concluded once eight replacement attempts (i.e., words for which no loss-reducing replacement has been found) have failed in a row.

Word-level search for bidirectional models

For models for which the evaluation of $\log p(s \mid m)$ is computationally cheap (2-gram, 3-gram, LSTM, and the RNN), we directly evaluated the log-probability of the 29,157 sentences resulting from each of the 29,157 possible word replacements. When such probability vectors were available for both models, we simply chose the replacement minimizing the loss. For GPT-2, whose evaluation is slower, we evaluated sentence probabilities only for word replacements for which the new word had a conditional log-probability (given the previous words in the sentence) of no less than −10; in rare cases when this threshold yielded fewer than 10 candidate words, we reduced the threshold in steps of 5 until there were at least 10 words above the threshold. For the bi-directional models (BERT, RoBERTa, XLM, and ELECTRA), for which the evaluation of $\log p(s \mid m)$ is costly even for a single sentence, we used a heuristic to prioritize which replacements to evaluate.

Since bi-directional models are trained as masked language models, they readily provide word-level completion probabilities. These word-level log-probabilities typically have positive but imperfect correlation with the log-probabilities of the sentences resulting from each potential completion. We hence formed a simple linear regression-based estimate of $\log p(s\{i\} \leftarrow w \mid m)$, the log-probability of the sentence s with word w assigned at position i, predicting it from $\log p(s\{i\} = w \mid m, s\{i\} \leftarrow mask)$, the completion log-probability of word w at position i, given the sentence with the i-th word masked (Eq. 3):

$$\log \hat{p}(s\{i\} \leftarrow w \mid m) = \beta_1 \log p(s\{i\} = w \mid m, s\{i\} \leftarrow mask) + \beta_0$$

This regression model was estimated from scratch for each word-level search. When a word was first selected for being replaced, the log-probability of two sentences was evaluated: the sentence resulting from substituting the existing word with the word with the highest completion probability and the sentence resulting from substituting the existing word with the word with the lowest completion probability. These two word-sentence log-probability pairs, as well as the word-sentence log-probability pair pertaining to the current word, were used to fit the regression line. The regression prediction, together with the sentence probability for the other model (either the exact probability, or approximate probability if the other model was also bidirectional) was used to predict $\log p(s \mid m_{reject})$ for each of the 29,157 potential replacements. We then evaluated the true (non-approximate) sentence probabilities of the replacement word with the minimal predicted probability. If this word indeed reduced the sentence probability, it was chosen to serve as the replacement and the word-level search was terminated (i.e., proceeding to search a replacement for another word in the sentence). If it did not reduce the probability, the regression model (Eq. 3) was updated with the new observation, and the next

replacement expected to minimize the sentence probability was evaluated. This word-level search was terminated after five sentence evaluations that did not reduce the loss.

Selecting the best sentence triplets from the optimization results

Since the discrete hill-climbing procedure described above is highly local, the degree to which this succeeded in producing highly-controversial pairs varied depending on the starting sentence *n*. We found that typically, natural sentences with lower than average log-probability gave rise to synthetic sentences with greater controversiality. To better represent the distribution of natural sentences while still choosing the best (most controversial) triplets for human testing, we used stratified selection.

First, we quantified the controversiality of each triplet as (Eq. 4):

$$c_{m1,m2}(n, s_1, s_2) = \log [p(n \mid m_1) / p(s_1 \mid m_1)] + \log [p(n \mid m_2) / p(s_2 \mid m_2)]$$

where s_1 is the sentence generated to reduce the probability in model m_1 and s_2 is the sentence generated to reduce the probability in model m_2.

We employed integer programming to choose the 10 most controversial triplets from the 100 triplets optimized for each model pair (maximizing the total controversiality across the selected triplets), while ensuring that for each model, there was exactly one natural sentence in each decile of the natural sentences probability distribution. The selected 10 synthetic triplets were then used to form 30 unique experimental trials per model pair, comparing the natural sentence with one synthetic sentence, comparing the natural sentence with the other synthetic sentence, and comparing the two synthetic sentences.

Design of the human experiment

Our experimental procedures were approved by the Columbia University Institutional Review Board (protocol number IRB-AAAS0252). All participants provided informed consent

prior. We presented the controversial sentence pairs selected and synthesized by the language models to 100 native English-speaking, US-based participants (55 male) recruited from Prolific (www.prolific.co), and paid each participant $5.95. The average participant age was 34.08 ± 12.32. The subjects were divided into 10 groups, and each ten-subject group was presented with a unique set of stimuli. Each stimulus set contained exactly one sentence pair from every possible combination of model pairs and the four main experimental conditions: selected controversial sentence pairs; natural vs. synthetic pair, where one model served as m_{accept} and the other as m_{reject}; a natural vs. synthetic pair with the reverse model role assignments; and directly pairing the two synthetic sentences. These model-pair-condition combinations accounted for 144 (36×4) trials of the task. In addition to these trials, each stimulus set also included nine trials consisting of sentence pairs randomly sampled from the database of eight-word sentences (and not already included in any of the other conditions). All subjects also viewed 12 control trials consisting of a randomly selected natural sentence and the same natural sentence with the words scrambled in a random order. The order of trials within each stimulus set as well as the left-right screen position of sentences in each sentence pair were randomized for all participants. While each sentence triplet produced by the optimization procedure (see subsection "Generating synthetic controversial sentence pairs") gave rise to three trials, these were allocated such that no subject viewed the same sentence twice.

On each trial of the task, participants were asked to make a binary decision about which of the two sentences they considered more probable (for the full set of instructions given to participants, see Fig. S1). In addition, they were asked to indicate one of three levels of confidence in their decision: somewhat confident, confident, or very confident. The trials were not timed, but a 90-minute time limit was enforced for the whole experiment. A progress bar at

the bottom of the screen indicated to participants how many trials they had completed and had remaining to complete.

We rejected the data of 21 participants who failed to choose the original, unshuffled sentence in at least 11 of the 12 control trials, and acquired data from 21 alternative participants instead, all of whom passed this data-quality threshold. In general, we observed high agreement in sentence preferences among our participants, though the level of agreement varied across conditions. There was complete or near-complete agreement (at least 9/10 participants with the same binary sentence preference) in 52.2% of trials for randomly-sampled natural-sentence pairs, 36.6% of trials for controversial natural-sentence pairs, 67.6% of trials for natural-synthetic pairs, and 60.0% of trials for synthetic-synthetic pairs (versus a chance rate of 1.1%, assuming a binomial distribution with $p = 0.5$).

Evaluation of model-human consistency

To measure the alignment on each trial between model judgments and human judgments, we binarized both measures; we determined which of the two sentences was assigned with a higher probability by the model, regardless of the magnitude of the probability difference, and which of the two sentences was favored by the subject, regardless of the reported confidence level. When both the subject and the model chose the same sentence, the trial was considered as correctly predicted by that model. This correctness measure was averaged across sentence pairs and across the 10 participants who viewed the same set of trials. For the lower bound on the noise ceiling, we predicted each subject's choices from a majority vote of the nine other subjects who were presented with the same trials. For the upper bound (i.e., the highest possible accuracy attainable on this data sample), we included the subject themselves in this majority vote-based prediction.

Since each of the 10 participant groups viewed a unique trial set, these groups provided 10 independent replications of the experiment. Models were compared to each other and to the lower bound of the noise ceiling by a Wilcoxon signed-rank test using these 10 independent accuracy outcomes as paired samples. For each analysis, the false discovery rate across multiple comparisons was controlled by the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995).

In Figure 1.4, we instead measure model-human consistency in a more continuous way, comparing the sentence probability ratio in a model to the graded Likert ratings provided by humans; see Supplementary Methods for full details.

Selecting trials for model evaluation

All of the randomly sampled natural-sentence pairs (Fig. 1.1a) were evaluated for each of the candidate models. Controversial sentence pairs (either natural, Fig. 1.1b or synthetic, Fig. 1.3) were included in a model's evaluation set only if they were formed to target that model specifically. The overall summary analysis (Fig. 1.4) evaluated all models on all available sentence pairs.

Comparison to an existing approach for computing sentence probabilities in bidirectional models

Wang & Cho (2019) have suggested an alternative approach for computing sentence probabilities in bidirectional (BERT-like) models, using a pseudo-log-likelihood measure which simply sums the log-probability of each token conditioned on all of the other tokens in the sentence. While this measure does not yield a probability measure Cho, 2019, it is positively correlated with human acceptability judgments for several bidirectional models (Lau et al., 2020; Salazar et al., 2020). To directly compare this existing approach to our novel method for

computing probabilities, we again used the method of controversial sentence pairs to identify the approach most aligned with human judgments. For each bidirectional model (BERT, RoBERTa, and ELECTRA), we created two copies of the model, each using a different approach for computing sentence probabilities. We synthesized 40 synthetic sentence pairs to maximally differentiate between the two copies of each model, with each copy assigning a higher probability to a different sentence in the pair. We then tested 30 human participants, presenting each participant with all 120 sentence pairs.

Data and code availability

Experimental stimuli, detailed behavioral testing results, sentence optimization code, and code for reproducing all analyses and figures are available at github.com/dpmlab/contstimlang.

1.3 Results

We acquired judgments from 100 native English speakers tested online. In each experimental trial, the participants were asked to judge which of two sentences they would be "more likely to encounter in the world, as either speech or written text", and provided a rating of their confidence in their answer on a 3-point scale (see Fig. S1 for task instructions and Fig. S2 for a trial example). The experiment was designed to compare nine different language models: probability models based on corpus frequencies of 2-word and 3-word sequences (2-grams and 3-grams) and a range of neural network models comprising a recurrent neural network (RNN), a long short-term memory network (LSTM), and five transformer models (BERT, RoBERTa, XLM, ELECTRA, and GPT-2).

Controversial natural-sentence pairs enable efficient model comparison

As a baseline, we randomly sampled and paired 8-word sentences from a corpus of Reddit comments. However, as shown in Figure 1.1a, these sentences fail to uncover meaningful

differences between the models. For each sentence pair, all models tend to prefer the same sentence, and therefore perform similarly in predicting human preference ratings (see Supplementary Results).

Instead, we can use an optimization procedure (Eq. 5, Methods) to search for controversial sentence pairs, in which one language model assigns a high probability (above the median probability for natural sentences) only to sentence 1 and a second language model assigns a high probability only to sentence 2; see examples in Table 1.1. Measuring each model's accuracy in predicting human choices for sentence pairs in which it was one of the two targeted models indicated many significant differences in terms of model-human alignment (Fig. 1.1b), with GPT-2 and RoBERTa showing the best human consistency and 2-gram the worst. We can also compare each model pair separately (using only the stimuli targeting that model pair), yielding a similar pattern of pairwise dominance (Fig. S4a). All models except GPT-2, RoBERTa, and ELECTRA performed significantly below our lower bound on the noise ceiling (the accuracy obtained by predicting each participant's responses from the other participants' responses), indicating a misalignment between these models' predictions and human judgments which was only revealed when using controversial sentence pairs.

sentence	log probability (model 1)	log probability (model 2)	# human choices
n_1: Rust is generally caused by salt and sand.	$\log p(n_1\|\text{GPT-2}) = -50.72$	$\log p(n_1\|\text{ELECTRA}) = -\mathbf{38.54}$	**10**
n_2: Where is Vernon Roche when you need him.	$\log p(n_2\|\text{GPT-2}) = -\mathbf{32.26}$	$\log p(n_2\|\text{ELECTRA}) = -58.26$	0
n_1: Excellent draw and an overall great smoking experience.	$\log p(n_1\|\text{RoBERTa}) = -67.78$	$\log p(n_1\|\text{GPT-2}) = -\mathbf{36.76}$	**10**
n_2: I should be higher and tied to inflation.	$\log p(n_2\|\text{RoBERTa}) = -\mathbf{54.61}$	$\log p(n_2\|\text{GPT-2}) = -50.31$	0
n_1: You may try and ask on their forum.	$\log p(n_1\|\text{ELECTRA}) = -51.44$	$\log p(n_1\|\text{LSTM}) = -\mathbf{44.24}$	**10**
n_2: I love how they look like octopus tentacles.	$\log p(n_2\|\text{ELECTRA}) = -\mathbf{35.51}$	$\log p(n_2\|\text{LSTM}) = -66.66$	0
n_1: Grow up and quit whining about minor inconveniences.	$\log p(n_1\|\text{BERT}) = -82.74$	$\log p(n_1\|\text{GPT-2}) = -\mathbf{35.66}$	**10**
n_2: The extra a is the correct Sanskrit pronunciation.	$\log p(n_2\|\text{BERT}) = -\mathbf{51.06}$	$\log p(n_2\|\text{GPT-2}) = -51.10$	0
n_1: I like my password manager for this reason.	$\log p(n_1\|\text{XLM}) = -68.93$	$\log p(n_1\|\text{RoBERTa}) = -\mathbf{49.61}$	**10**
n_2: Kind of like clan of the cave bear.	$\log p(n_2\|\text{XLM}) = -\mathbf{44.24}$	$\log p(n_2\|\text{RoBERTa}) = -67.00$	0
n_1: We have raised a Generation of Computer geeks.	$\log p(n_1\|\text{LSTM}) = -66.41$	$\log p(n_1\|\text{ELECTRA}) = -\mathbf{36.57}$	**10**

n_2: I mean when the refs are being sketchy.	log$p$(n_2|LSTM) =−**42.04**	log$p$(n_2|ELECTRA) =−52.28	0
n_1: This is getting ridiculous and ruining the hobby.	log$p$(n_1|RNN) =−100.65	log$p$(n_1|LSTM) =−**43.50**	**10**
n_2: I think the boys and invincible are better.	log$p$(n_2|RNN) =−**45.16**	log$p$(n_2|LSTM) =−59.00	0
n_1: Then attach them with the supplied wood screws.	log$p$(n_1|3-gram) =−119.09	log$p$(n_1|GPT-2) =−**34.84**	**10**
n_2: Sounds like you were used both a dog.	log$p$(n_2|3-gram) =−**92.07**	log$p$(n_2|GPT-2) =−52.84	0
n_1: Cream cheese with ham and onions on crackers.	log$p$(n_1|2-gram) =−131.99	log$p$(n_1|RoBERTa) =−**54.62**	**10**
n_2: I may have to parallel process that drinking.	log$p$(n_2|2-gram) =−**109.46**	log$p$(n_2|RoBERTa) =−70.69	0

Table 1.1: Examples of controversial natural-sentence pairs that maximally contributed to each model's prediction error. For each model (double row, "model 1"), the table shows results for two sentences on which the model failed severely. In each case, the failing model 1 prefers sentence n_2 (higher log probability bolded), while the model it was pitted against ("model 2") and all 10 human subjects presented with that sentence pair prefer sentence n_1. (When more than one sentence pair induced an equal maximal error in a model, the example included in the table was chosen at random.)

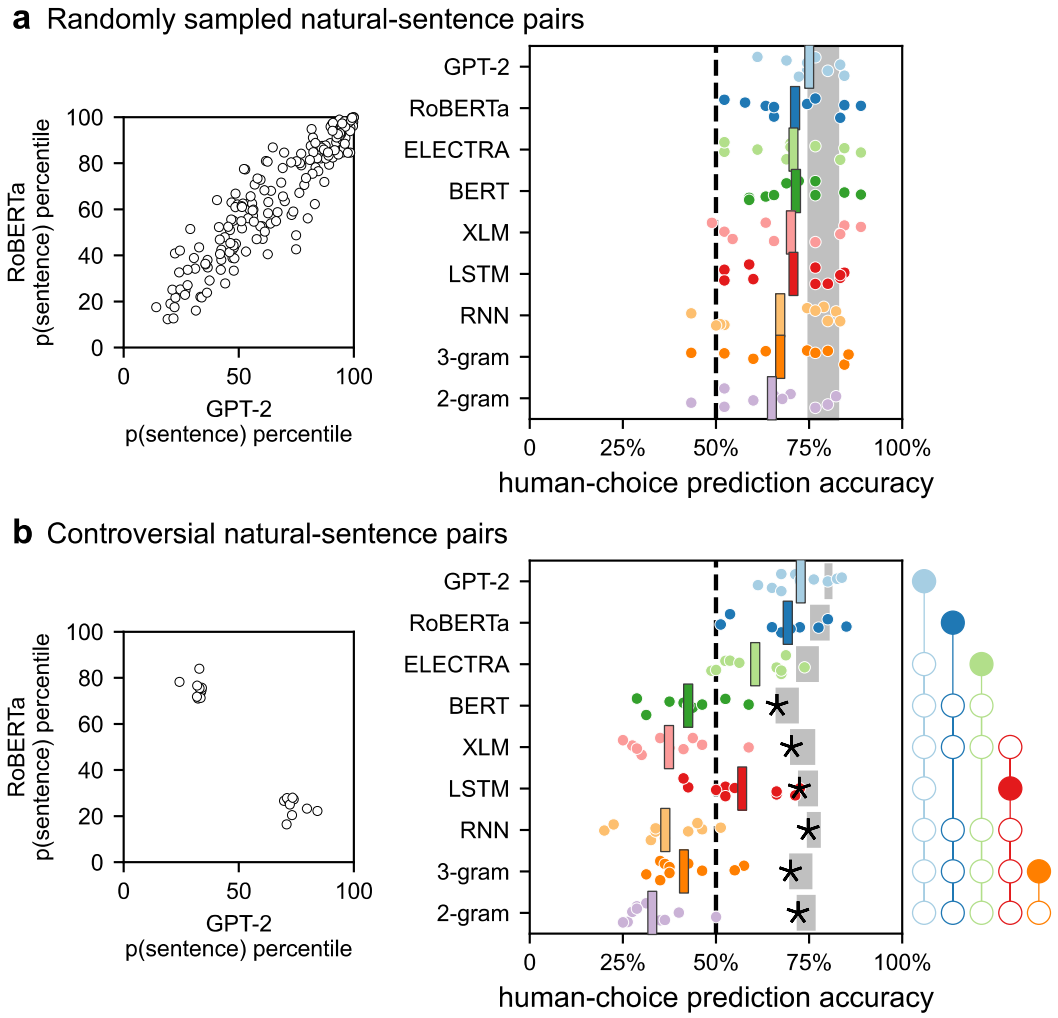

Figure 1.1: Model comparison using natural sentences. (a) (Left) Percentile-transformed sentence probabilities for GPT-2 and RoBERTa (defined relative to all sentences used in the experiment) for randomlysampled pairs of natural sentences. Each pair of connected dots depicts one sentence pair. The two models are highly congruent in their rankings of sentences within a pair (lines have upward slope). (Right) Accuracy of model predictions of human choices, measured as the proportion of trials in which the same sentence was preferred by both the model and the human participant. Each dot depicts the prediction accuracy of one candidate model averaged across a group of 10 participants presented with a unique set of trials. The colored bars depict grand-averages across all 100 participants. The gray bar is the noise ceiling whose left and right edges are lower and upper bounds on the grand-average performance an ideal model would achieve (based on the consistency across human subjects). There were no significant differences in model performance on the randomly sampled natural sentences**. (b)** (Left) Controversial natural-sentence pairs were selected such that the models' sentence probability ranks were incongruent (lines have downward slope). (Right) Controversial sentence pairs enable efficient model comparison, revealing that BERT, XLM, LSTM, RNN and the n-gram models perform significantly below the noise ceiling (asterisks indicate significance–two-sided Wilcoxon signed-rank test, controlling the false discovery rate for nine comparisons at $q < .05$). On the right of the plot, each closed circle indicates a model significantly dominating alternative models indicated by open circles (two-sided Wilcoxon signed-rank test, controlling the false discovery rate for all 36 model pairs at $q < .05$). GPT-2 outperforms all models except RoBERTA at predicting human judgments.

Synthesized controversial sentence pairs enable even greater disentanglement of model predictions

Selecting controversial natural-sentence pairs may provide greater power than randomly sampling natural-sentence pairs, but this search procedure considers a very limited part of the space of possible sentence pairs. Instead, we can iteratively replace words in a natural sentence to drive different models to make opposing predictions, forming synthetic controversial sentences that may lay outside any natural language corpora, as illustrated in Figure 1.2 (see Methods, "Generating synthetic controversial sentence pairs" for full details). Examples of controversial synthetic-sentence pairs that maximally contributed to the models' prediction error appear in Table 1.2.

We evaluated how well each model predicted the human sentence choices in all of the controversial synthetic-sentence pairs in which the model was one of the two models targeted

(Fig. 1.3a). This evaluation of model-human alignment resulted in an even greater separation between the models' prediction accuracies than was obtained when using controversial natural sentence pairs, pushing the weaker models (RNN, 3-gram, and 2-gram) far below the 50%

Figure 1.2: Synthesizing controversial sentence pairs. The small open dots denote 500 randomly sampled natural sentences. The big open dot denotes the natural sentence used for initializing the controversial sentence optimization, and the closed dots are the resulting synthetic sentences. (a) In this example, we start with the randomly sampled natural sentence "Luke has a ton of experience with winning". If we adjust this sentence to minimize its probability according to GPT-2 (while keeping the sentence at least as likely as the natural sentence according to ELECTRA), we obtain the synthetic sentence "Nothing has a world of excitement and joys". By repeating this procedure while switching the roles of the models, we generate the synthetic sentence "Diddy has a wealth of experience with grappling", which decreases ELECTRA's probability while slightly increasing GPT-2's. (b) In this example, we start with the randomly sampled natural sentence "I need to see how this played out". If we adjust this sentence to minimize its probability according to RoBERTa (while keeping the sentence at least as likely as the natural sentence according to 3-gram), we obtain the synthetic sentence "You have to realize is that noise again". If we instead decrease only 3-gram's probability, we generate the synthetic sentence "I wait to see how it shakes out".

sentence	log probability (model 1)	log probability (model 2)	# human choices
s_1: You can reach his stories on an instant.	$\log p(s_1\|\text{GPT-2}) = -64.92$	$\log p(s_1\|\text{RoBERTa}) = -\mathbf{59.98}$	**10**
s_2: Anybody can behead a rattles an an antelope.	$\log p(s_2\|\text{GPT-2}) = -\mathbf{40.45}$	$\log p(s_2\|\text{RoBERTa}) = -90.87$	0
s_1: However they will still compare you to others.	$\log p(s_1\|\text{RoBERTa}) = -53.40$	$\log p(s_1\|\text{GPT-2}) = -\mathbf{31.59}$	**10**
s_2: Why people who only give themselves to others.	$\log p(s_2\|\text{RoBERTa}) = -\mathbf{48.66}$	$\log p(s_2\|\text{GPT-2}) = -47.13$	0
s_1: He healed faster than any professional sports player.	$\log p(s_1\|\text{ELECTRA}) = -48.77$	$\log p(s_1\|\text{BERT}) = -\mathbf{50.21}$	**10**
s_2: One gets less than a single soccer team.	$\log p(s_2\|\text{ELECTRA}) = -\mathbf{38.25}$	$\log p(s_2\|\text{BERT}) = -59.09$	0

s_1: That is the narrative we have been sold.	$\log p(s_1\|\text{BERT}) = -56.14$	$\log p(s_1\|\text{GPT-2}) = \mathbf{-26.31}$	**10**
s_2: This is the week you have been dying.	$\log p(s_2\|\text{BERT}) = \mathbf{-50.66}$	$\log p(s_2\|\text{GPT-2}) = -39.50$	0
s_1: The resilience is made stronger by early adversity.	$\log p(s_1\|\text{XLM}) = -62.95$	$\log p(s_1\|\text{RoBERTa}) = \mathbf{-54.34}$	**10**
s_2: Every thing is made alive by infinite Ness.	$\log p(s_2\|\text{XLM}) = \mathbf{-42.95}$	$\log p(s_2\|\text{RoBERTa}) = -75.72$	0
s_1: President Trump threatens to storm the White House.	$\log p(s_1\|\text{LSTM}) = -58.78$	$\log p(s_1\|\text{RoBERTa}) = \mathbf{-41.67}$	**10**
s_2: West Surrey refused to form the White House.	$\log p(s_2\|\text{LSTM}) = \mathbf{-40.35}$	$\log p(s_2\|\text{RoBERTa}) = -67.32$	0
s_1: Las beans taste best with a mustard sauce.	$\log p(s_1\|\text{RNN}) = -131.62$	$\log p(s_1\|\text{RoBERTa}) = \mathbf{-60.58}$	**10**
s_2: Roughly lanes being alive in a statement ratings.	$\log p(s_2\|\text{RNN}) = \mathbf{-49.31}$	$\log p(s_2\|\text{RoBERTa}) = -99.90$	0
s_1: You are constantly seeing people play the multi.	$\log p(s_1\|\text{3-gram}) = -107.16$	$\log p(s_1\|\text{ELECTRA}) = \mathbf{-44.79}$	**10**
s_2: This will probably the happiest contradicts the hypocrite.	$\log p(s_2\|\text{3-gram}) = \mathbf{-91.59}$	$\log p(s_2\|\text{ELECTRA}) = -75.83$	0
s_1: A buyer can own a genuine product also.	$\log p(s_1\|\text{2-gram}) = -127.35$	$\log p(s_1\|\text{ELECTRA}) = \mathbf{-40.21}$	**10**
s_2: One versed in circumference of highschool I rambled.	$\log p(s_2\|\text{2-gram}) = \mathbf{-113.73}$	$\log p(s_2\|\text{ELECTRA}) = -92.61$	0

Table 1.2: Examples of controversial synthetic-sentence pairs that maximally contributed to each model's prediction error. For each model (double row, "model 1"), the table shows results for two sentences on which the model failed severely. In each case, the failing model 1 prefers sentence s2 (higher log probability bolded), while the model it was pitted against ("model 2") and all 10 human subjects presented with that sentence pair prefer sentence s1. (When more than one sentence pair induced an equal maximal error in a model, the example included in the table was chosen at random.)

chance accuracy level. GPT-2, RoBERTa and ELECTRA were found to be significantly more accurate than the alternative models (BERT, XLM, LSTM, RNN, 3-gram, and 2-gram) in predicting the human responses to these trials (with similar results when comparing model pair separately, see Fig. S4b). All of the models except for GPT-2 were found to be significantly below the lower bound on the noise ceiling, demonstrating misalignment with human judgments.

Pairs of natural and synthetic sentences uncover blindspots in all models

Last, we considered trials in which the participants were asked to choose between a natural sentence and one of the synthetic sentences which was generated from that natural sentence. If the language model is fully aligned with human judgments, we would expect humans to agree with the model, and select the synthetic sentence at least as much as the natural sentence. In reality, human participants showed a systematic preference for the natural sentences over their synthetic counterparts (Fig. 1.3b), even when the synthetic sentences were formed such that the stronger models (i.e., GPT-2, RoBERTA, or ELECTRA) favored them over the

natural sentences; see Table 1.3 for examples. Evaluating natural sentence preference separately for each model-pairing (Fig. S5), we find that these imperfections can be uncovered even when pairing a strong model with a relatively weak model (such that the strong model "accepts" the synthetic sentence and the weak model rejects it).

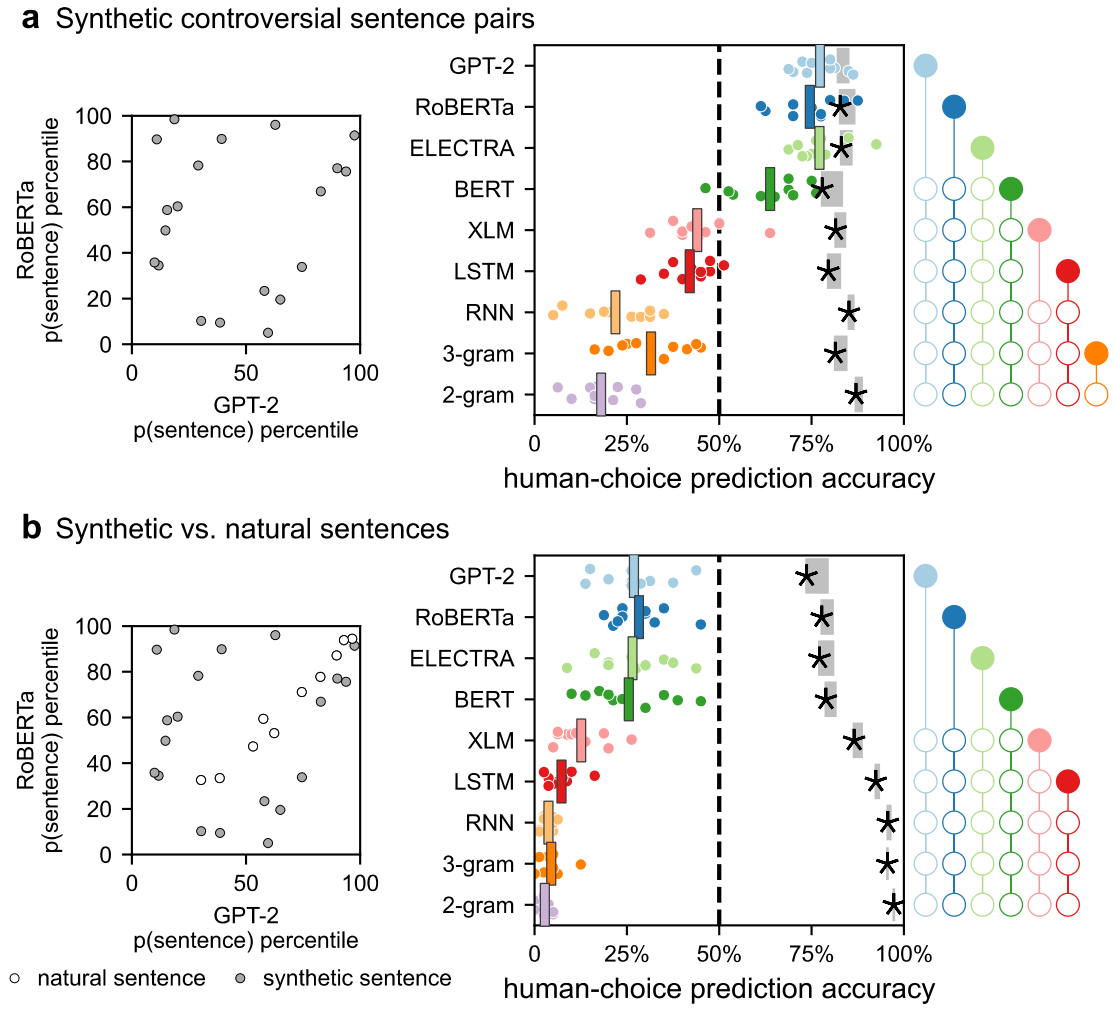

Figure 1.3: Model comparison using synthetic sentences. (a) (Left) Percentile-transformed sentence probabilities for GPT-2 and RoBERTa for controversial synthetic-sentence pairs. Each pair of connected dots depict one sentence pair. (Right) Model prediction accuracy, measured as the proportion of trials in which the same sentence was preferred by both the model and the human participant. GPT-2, RoBERTa and ELECTRA significantly outperformed the other models (two-sided Wilcoxon signed-rank test, controlling the false discovery rate for all 36 model comparisons at $q < .05$). All of the models except for GPT-2 were found to perform below the noise ceiling (gray) of predicting each participant's choices from the majority votes of the

other participants (asterisks indicate significance–two-sided Wilcoxon signed-rank test, controlling the false discovery rate for nine comparisons at $q < .05$). **(b)** (Left) Each connected triplet of dots depicts a natural sentence and its derived synthetic sentences, optimized to decrease the probability only under GPT-2 (left dots in a triplet) or only under RoBERTa (bottom dots in a triplet). (Right) Each model was evaluated across all of the synthetic-natural sentence pairs for which it was targeted to keep the synthetic sentence at least as probable as the natural sentence (see Fig. S6 for the complementary data binning). This evaluation yielded a below-chance prediction accuracy for all of the models, which was also significantly below the lower bound on the noise ceiling. This indicates that, although the models assessed that these synthetic sentences were at least as probable as the original natural sentence, humans disagreed and showed a systematic preference for the natural sentence. See Fig. 1.1's caption for details on the visualization conventions used in this figure.

sentence	log probability (model 1)	log probability (model 2)	# human choices
n: I always cover for him and make excuses.	log*p*(*n*|GPT-2) =−36.46	log*p*(*n*|2-gram) =−**106.95**	**10**
s: We either wish for it or ourselves do.	log*p*(*s*|GPT-2) =−**36.15**	log*p*(*s*|2-gram) =−122.28	0
n: This is why I will never understand boys.	log*p*(*n*|RoBERTa) =−46.88	log*p*(*n*|2-gram) =−**103.11**	**10**
s: This is why I will never kiss boys.	log*p*(*s*|RoBERTa) =−**46.75**	log*p*(*s*|2-gram) =−107.91	0
n: One of the ones I did required it.	log*p*(*n*|ELECTRA) =−35.97	log*p*(*n*|LSTM) =−**40.89**	**10**
s: Many of the years I did done so.	log*p*(*s*|ELECTRA) =−**35.77**	log*p*(*s*|LSTM) =−46.25	0
n: There were no guns in the Bronze Age.	log*p*(*n*|BERT) =−48.48	log*p*(*n*|ELECTRA) =−**30.40**	**10**
s: There is rich finds from the Bronze Age.	log*p*(*s*|BERT) =−**48.46**	log*p*(*s*|ELECTRA) =−44.34	0
n: You did a great job on cleaning them.	log*p*(*n*|XLM) =−40.38	log*p*(*n*|RNN) =−**43.47**	**10**
s: She did a great job at do me.	log*p*(*s*|XLM) =−**39.89**	log*p*(*s*|RNN) =−61.03	0
n: This logic has always seemed flawed to me.	log*p*(*n*|LSTM) =−39.77	log*p*(*n*|RNN) =−**45.92**	**10**
s: His cell has always seemed instinctively to me.	log*p*(*s*|LSTM) =−**38.89**	log*p*(*s*|RNN) =−62.81	0
s: Stand near the cafe and sip your coffee.	log*p*(*s*|RNN) =−65.55	log*p*(*s*|ELECTRA) =−**34.46**	**10**
n: Sit at the front and break your neck.	log*p*(*n*|RNN) =−**44.18**	log*p*(*n*|ELECTRA) =−34.65	0
n: Most of my jobs have been like this.	log*p*(*n*|3-gram) =−80.72	log*p*(*n*|LSTM) =−**35.07**	**10**
s: One of my boyfriend have been like this.	log*p*(*s*|3-gram) =−**80.63**	log*p*(*s*|LSTM) =−41.44	0
n: They even mentioned that I offer white flowers.	log*p*(*n*|2-gram) =−113.38	log*p*(*n*|BERT) =−**62.81**	**10**
s: But even fancied that would logically contradictory philosophies.	log*p*(*s*|2-gram) =−**113.24**	log*p*(*s*|BERT) =−117.98	0

Table 1.3: Examples of pairs of synthetic and natural sentences that maximally contributed to each model's prediction error. For each model (double row, "model 1"), the table shows results for two sentences on which the model failed severely. In each case, the failing model 1 prefers synthetic sentence *s* (higher log probability bolded), while the model it was pitted against ("model 2") and all 10 human subjects presented with that sentence pair prefer natural sentence *n*. (When more than one sentence pair induced an equal maximal error in a model, the example included in the table was chosen at random.)

Evaluating the entire dataset reveals a hierarchy of language models, but no model is fully human aligned

Rather than evaluating each model's prediction accuracy with respect to the particular sentence pairs that were formed to compare this model to alternative models, we can maximize our statistical power by computing the average prediction accuracy for each model with respect to all of the experimental trials we collected. Furthermore, rather than binarizing the human and model judgments, here we measure the ordinal correspondence between the graded human choices (taking confidence into account) and the log ratio of the sentence probabilities assigned by each candidate model. Using this more sensitive benchmark (Fig. 1.4), we found GPT-2 to be the most human-aligned, followed by RoBERTa; then ELECTRA; BERT; XLM and LSTM; and the RNN, 3-gram, and 2-gram models. However, all of the models (including GPT-2) were found to be significantly less accurate than the lower bound on the noise ceiling. One possible reason for the poorer performance of the bidirectional transformers (RoBERTa, ELECTRA, BERT, and XLM) compared to the unidirectional transformer (GPT-2) is that computing sentence probabilities in these models is complex, and the probability estimator we developed (see Methods, "Evaluating sentence-level probabilities in bidirectional models") could be non-optimal; however, when directly comparing our estimator to the popular pseudo-log-likelihood approach by means of synthetic controversial sentences, our estimator was found to be better aligned to human judgments (see Fig. S8 and Supplementary Results).

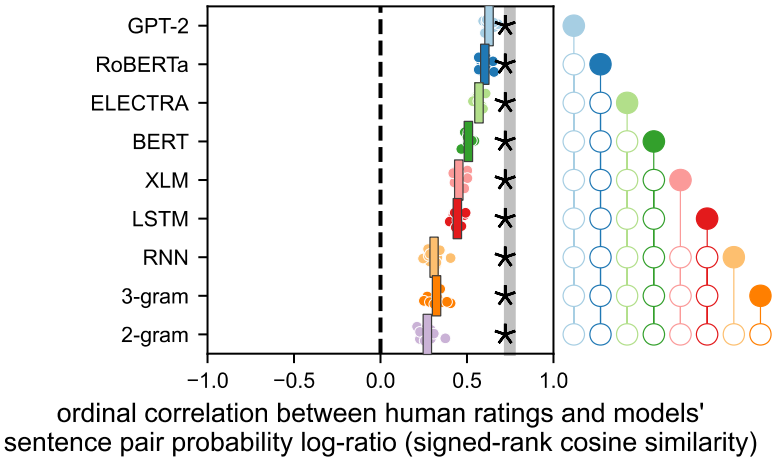

Figure 1.4: Ordinal correlation of the models' sentence probability log-ratios and human Likert ratings. For each sentence pair, model prediction was quantified by log [$p(s^1|m)$ / $p(s^2|m)$]. This log-ratio was correlated with the Likert ratings of each particular participant, using signed-rank cosine similarity (see Methods). This analysis, taking all trials and human confidence level into account, indicates that GPT-2 performed best in predicting human sentence probability judgments. However, its predictions are still significantly misaligned with the human choices. See Fig. 1.1's caption for details on the visualization convention.

1.4 Discussion

In this study, we probed the ability of language models to predict human relative sentence probability judgments using controversial sentence pairs, selected or synthesized so that two models disagreed about which sentence was more probable. We found that 1) GPT-2 (a unidirectional transformer model trained on predicting upcoming tokens) and RoBERTa (a bidirectional transformer trained on a held-out token prediction task) were the most predictive of human judgments on controversial natural-sentence pairs (Fig. 1.1b); 2) GPT-2, RoBERTa, and ELECTRA (a bidirectional transformer trained on detecting corrupted tokens) were the most predictive of human judgments on pairs of sentences synthesized to maximize controversiality (Fig. 1.3a); and 3) GPT-2 was the most human-consistent model when considering the entire behavioral dataset we collected (Fig. 1.4). And yet, all of the models, including GPT-2, exhibited behavior inconsistent with human judgments; using an alternative model as a counterforce, we

could corrupt natural sentences such that their probability under a model did not decrease, but humans tended to reject the corrupted sentence as unlikely (Fig. 1.3b).

Implications for artificial neural network language models as neuropsycholinguistic models

Unlike convolutional neural networks, whose architectural design principles are roughly inspired by biological vision (Lindsay, 2021), the design of current neural network language models is largely uninformed by psycholinguistics and neuroscience. And yet, there is an ongoing effort to adopt and adapt neural network language models to serve as computational hypotheses of how humans process language, making use of a variety of different architectures, training corpora, and training tasks (Wehbe et al., 2014; Toneva & Wehbe, 2019; Heilbron et al., 2020; Jain et al., 2020; Lyu et al., 2021; Schrimpf et al., 2021; Wilcox et al., 2021; Goldstein et al., 2022; Caucheteux & King, 2022; Arehalli et al., 2022). We found that recurrent neural networks make markedly human-inconsistent predictions once pitted against transformer-based neural networks. This finding coincides with recent evidence that transformers also outperform recurrent networks for predicting neural responses as measured by ECoG or fMRI (Schrimpf et al., 2021; Goldstein et al., 2022), as well as with evidence from model-based prediction of human reading speed (Wilcox et al., 2021; Merkx & Frank, 2021) and N400 amplitude (Merkx & Frank, 2021; Michaelov et al., 2021). Among the transformers, GPT-2, RoBERTa, and ELECTRA showed the best performance. These models are trained to optimize only word-level prediction tasks, as opposed to BERT and XLM which are additionally trained on next-sentence prediction and cross-lingual tasks, respectively (and have the same architecture as RoBERTa). This suggests that local word prediction provides better alignment with human language comprehension.

Despite the agreement between our results and previous work in terms of model ranking, the significant failure of GPT-2 in predicting the human responses to natural versus synthetic controversial pairs (Fig. 1.3b) demonstrates that GPT-2 does not fully emulate the computations employed in human processing of even short sentences. This outcome is in some ways unsurprising, given that GPT-2 (like all of the other models we considered) is an off-the-shelf machine learning model that was not designed with human psycholinguistic and physiological details in mind. And yet, the considerable human inconsistency we observed seems to stand in stark contrast with the recent report of GPT-2 explaining about 100 percent of the explainable variance in fMRI and ECoG responses to natural sentences (Schrimpf et al., 2021). Part of this discrepancy could be explained by the fact that Schrimpf and colleagues (Schrimpf et al., 2021) mapped GPT-2 hidden-layer activations to brain data by means of regularized linear regression, which can identify a subspace within GPT-2's language representation that is wellaligned with brain responses even if GPT-2's overall sentence probabilities are not human-like. More importantly, when language models are evaluated with natural language, strong statistical models might capitalize on features in the data that are distinct from, but highly correlated with, features that are meaningful to humans. Therefore, a model that performs well on typical sentences might employ computational mechanisms that are very distinct from the brain's, which will only be revealed by testing the model in a more challenging domain. Note that even the simplest model we considered—a 2-gram frequency table—actually performed quite well on predicting human judgments for randomly-sampled natural sentences, and its deficiencies only became obvious when challenged by controversial sentence pairs. We predict that there will be substantial discrepancies between neural representations and current language models when using stimuli that intentionally stress-test this relationship, using our proposed sentence-level controversiality

approach or complementary ideas such as maximizing controversial transition probabilities between consecutive words (Rakocevic, 2021).

Using controversial sentences can be seen as a generalization test of language models: can models predict what kinds of changes to a natural sentence will lead to humans rejecting the sentence as improbable? Humans are sometimes capable of comprehending language with atypical constructions (e.g. in cases when pragmatic judgments can be made about a speaker's intentions from environmental and linguistic context, Goodman & Frank, 2016), but none of the models we tested were fully able to predict which syntactic or semantic perturbations would be accepted or rejected by humans. One possibility is that stronger next-word prediction models, using different architectures, learning rules, or training data, might close the gap between models and humans. Alternatively, it might be that optimizing for other linguistic tasks, or even non-linguistic task demands (in particular, representing the external world, the self, and other agents) will turn out to be critical for achieving human-like natural language processing (Howell et al., 2005).

Pitting models against each other circumvents the ground-truth problem of adversarial methods for language models

Machine vision models are highly susceptible to adversarial examples (Szegedy et al., 2013; Goodfellow et al., 2015). Such adversarial examples are typically generated by choosing a correctly classified natural image and then searching for a minuscule (and therefore human-imperceptible) image perturbation that would change the targeted model's classification. The prospect that similar covert model failure modes may exist also for language models has motivated proposed generalizations of adversarial methods to textual inputs (Zhang et al., 2020). However, imperceptible perturbations cannot be applied to written text: any modified word or

character is humanly perceptible. Prior work on adversarial examples for language models have instead relied on heuristic constraints aiming to limit the change in the meaning of the text, such as flipping a character (Liang et al., 2018; Ebrahimi et al., 2018), changing number or gender (Abdou et al., 2020), or replacing words with synonyms (Alzantot et al., 2018; Ribeiro et al., 2018; Ren et al., 2019). However, since these heuristics are only rough approximations of human language processing, many of these methods fail to preserve semantic meaning (Morris et al., 2020). Interactive ("human-in-the-loop") adversarial approaches allow human subjects to repeatedly alter model inputs such that it confuses target models but not secondary participants (Wallace et al., 2019; Kiela et al., 2021), but these approaches are inherently slow and costly and are limited by mental models the human subjects form about the evaluated language models.

By contrast, testing language models on controversial sentence pairs does not require approximating or querying a human ground truth during optimization–the objective of controversiality is independent of correctness. Instead, by designing inputs to elicit conflicting predictions among the models and assessing human responses to these inputs only once the optimization loop has terminated, we capitalize on the simple fact that if two models disagree with respect to an input, at least one of the models must be making an incorrect prediction. Pitting language models against other language models also can be conducted by other approaches such as "red-teaming", where an alternative language model is used as a generator of potential adversarial examples for a targeted model and a classifier is used to filter the generated examples such that the output they induce in the targeted model is indeed incorrect (Perez et al., 2022). Our approach shares the underlying principle that an alternative language model can drive a more powerful test than handcrafted heuristics, but here the models have symmetric roles

(there are no “attacking” and “attacked” models) and we can optimize stimuli directly without relying on filtering.

Limitations and future directions

While our results demonstrate that using controversial stimuli can identify subtle differences in language models’ alignment with human judgments, our study was limited in a number of ways. Our stimuli were all 8-word English sentences, limiting our ability to make cognitively meaningful claims that apply to language use globally. 8-word sentences are long enough to include common syntactic constructions and convey meaningful ideas but may not effectively probe long-distance syntactic dependencies (Gibson, 1998). Future work may introduce additional sentence lengths and languages, as well as (potentially adaptive) controversial sentence optimization procedures that consider large sets of candidate models, allowing for greater model coverage than our simpler pairwise approach. Future work may also complement the model-comparative experimental design with procedures designed to identify potential failure modes common to all models.

A more substantial limitation of the current study is that, like any comparison of pre-trained neural networks as potential models of human cognition, there could be multiple reasons (i.e., training data, architecture, training tasks, and learning rules) why particular models are better aligned with human judgments. For example, as we did not systematically control the training corpora used for training the models, it is possible that some of the observed differences are due to differences in the training sets rather than model architecture. Therefore, while our results expose failed model predictions, they do not readily answer why these failed predictions arise. Future experiments could compare custom-trained or systematically manipulated models, which reflect specific hypotheses about human language processing. In Figure S8, we

demonstrate the power of using synthetic controversial stimuli to conduct sensitive comparisons between models with subtle differences in how sentence probabilities are calculated.

It is important to note that our analyses considered human relative probability judgments as reflecting a scalar measure of acceptability. We made this assumption in order to bring the language models (which assign a probability measure to each sentence) and the human participants onto a common footing. However, it is possible that different types of sentence pairs engage different human cognitive processes. For pairs of synthetic sentences, both sentences may be unacceptable in different ways (e.g. exhibit different kinds of grammatical violations), requiring a judgment that weighs the relative importance of multiple dimensions (Watt, 1975) and could therefore produce inconsistent rankings across participants or across trials (Schutze, 2016). By contrast, asking participants to compare a natural and a synthetic sentence (Fig. 1.3b, Table 1.3) may be more analogous to previous work measuring human acceptability judgments for sentence pairs (Schutze & Sprouse, 2014). Nonetheless, it is worth noting that for all of the controversial conditions, the noise ceiling was significantly above the models' prediction accuracy, indicating non-random human preferences unexplained by current models that should be accounted for by future models, which may have to be more complex and capture multiple processes.

Finally, the use of synthetic controversial sentences can be extended beyond probability judgments. A sufficiently strong language model may enable constraining the experimental design search-space to particular sentence distributions (e.g., movie reviews or medical questions). Given such a constrained space, we may be able to search for well-formed sentences that elicit contradictory predictions in alternative domain-specific models (e.g., sentiment

classifiers or question-answering models). However, as indicated by our results, the task of capturing distributions of well-formed sentences is less trivial than it seems.

Chapter 2: Investigating naturalistic schema learning with computer-generated poetry

2.1 Introduction

Much has been learned in recent years about the cognitive architecture supporting the perception of naturalistic event scripts – stereotypical temporal sequences that reflect the shared structure of a set of related events. Important discoveries have been made about the neural mechanisms, temporal dynamics and semantic representations that define this category of experience. Numerous studies have shown that continuous naturalistic experiences are marked by *event boundaries* – shifts between stable patterns of neural activity at event transitions – which facilitate the mental segmentation of experiences in perception and memory (Kurby & Zacks, 2008). While event segmentation has long been considered an internally valid cognitive function that can be reliably measured in behavior (Newtson, 1976), there has been disagreement in the literature as to what exactly triggers event boundaries in the brain. The most well-known "event segmentation theory" holds that segmentation is a response to implicit prediction error between bottom-up sensory inputs and top-down "working event models" (Zacks et al., 2007; Zacks et al., 2009), but recent work has shown that even predictable environmental changes are similarly linked to ongoing segmentation (Clewett et al., 2019). For example, event boundaries have been measured in response to repeated viewings of the same stimulus (Lee et al., 2021) and to expected switches between task demands (Wang & Egner, 2022).

This disagreement may stem from the complex hierarchy of timescales, concepts, and brain regions in which event boundaries have been identified (Kuperberg, 2021), with more transient environmental changes corresponding to shorter events in low-level sensory brain areas

such as the early visual system, and deeper transitions between longer meaningful experiences corresponding to high-order areas of the default mode network (DMN), including posterior medial cortex (PMC) and the medial prefrontal cortex (mPFC) (Baldassano et al., 2017; Baldassano et al., 2018). Furthermore, this hierarchical relationship between environmental inputs and neural responses has been shown to depend on prior knowledge and inferences (Shin & DuBrow, 2021). Therefore, a full accounting of what constitutes an event boundary depends *at least* on the brain region in which neural activity is measured, the stability of features in the environment that are represented or processed by that brain region, and the access of that brain region to stored knowledge that is predictive of those features.

Naturalistic event schemas likely play a critical role in how these complex relationships play out ecologically. For example, when navigating an airport, a high-level model of event sequences (entering the airport, passing through security, waiting at the gate, boarding an airplane) may be active in mPFC, while an intermediate model of event content (announcements, luggage, terminal halls) may be active in PMC, and a perceptual model of low-level features (sounds, shapes) may be active in early sensory areas (Hasson et al., 2015; Baldassano et al., 2017; Baldassano et al., 2018). Naturalistic event schemas in this sense could be considered experiences whose segmentation in the brain is predicted in part by prior semantic knowledge, as opposed to superficial stimulus associations or task demands.

In this current work, we investigate how such schemas are *learned* and the *impact* of this learning on event representations. We created a novel event schema by embedding ten common semantic *topics* into four-line, iambic pentameter stanzas of rhyming poetry, and then ordered the topics into two arbitrary sequences (Sequence One and Sequence Two). By measuring the brain activity of 32 participants (16 per sequence) in fMRI listening to spoken recordings of the

poetry before and after learning its topic sequence, we aimed to describe how the segmentation and representation of schematic content *change* across the cortex as a result of learning, particularly in regions of the DMN. Comparing neural responses to schema-consistent stimuli in the presence and absence of prior knowledge could yield cognitive markers of an active schema model in a given brain area, which in turn could help determine whether previous findings on event segmentation are necessarily linked to schema knowledge, potentially contributing to the current debate surrounding event boundaries more generally. Furthermore, we aim to discover how the semantic content of discrete events becomes represented and temporally associated across the DMN. Understanding which key brain regions are flexible with respect to the semantic features they represent in and out of schematic contexts would highlight regions of interest for future work.

We took a unique approach by using a computer poetry generator to write the stimuli for this experiment. The generator utilized two different models: i) BERT (Devlin et al., 2019) is a bidirectional transformer model that outputs probability distributions for held-out words based on left and rightward context; ii) GloVe (Pennington et al., 2014) is a model of semantic word vectors learned by counting word co-occurrences in large text corpora, which has been previously used to decode semantic representations in fMRI data in response to a wide variety of concrete and abstract topics (Pereira et al., 2018). We used the GloVe model to define the topics in the schema, and then to constrain the output from BERT so that its content was tied to a topic-dependent distribution of features in GloVe vector space (see Fig. 2.1). The purpose of this method was three-fold: i) By generating hundreds of thousands of stanzas related to the ten core topics, we had sufficient material to design fMRI tasks that included over 60 total minutes of unique, schema-consistent poetry, and a schema learning task that presented additional unique

stanzas randomly selected from a large corpus of poetry. ii) By ensuring that the poetry heard before and after schema learning was generated using the same procedure, we could precisely match the poems' semantic content between conditions, allowing us to isolate learning-specific effects in fMRI data. iii) By constraining the semantic space of the topics in the poetry with the GloVe model, we were able to then use the GloVe model to make predictions about fMRI responses to the topics.

Here, we report the progress that has been made so far on a series of related fMRI analyses. A searchlight analysis tested correlations in the time course activity of individual five-stanza poems measured before and after learning (by different groups of participants) and found areas in PMC and mPFC where participants showed consistent learning-induced changes. A subsequent Hidden Markov Model (HMM) analysis of event boundary strength highlighted overlapping brain areas with a small increase in boundary magnitude after schema learning. Another analysis fit a general linear model to fMRI time course data to isolate topic coefficients predictive of voxelwise activity and found consistent learning-induced changes in the spatial correlations of these values in the PMC. A subsequent representational similarity analysis (RSA) found that changes to the topic coefficients can be meaningfully described by their similarity to the spatial geometry of the topics in the GloVe model.

In the process of collecting data for this experiment, the first 16 participants were administered tasks with the schema ordered in Sequence One. Their fMRI data was then analyzed. Then, two years later, the latter 16 participants were administered the same tasks with Sequence Two. In this sense, we performed a self-replication of this experiment, and additionally tested whether our initial results were robust to the order of topics in the schema. We found that the general topography of learning-induced changes to both poem time courses and topic

representations identified in Sequence One participants was conserved in multiple regions of the DMN in Sequence Two participants. In the two model-based analyses – the HMM and RSA analyses – a different profile of post-learning changes was observed between the two sequences. We report maps displaying the results of these analyses along the medial cortical surface in each group of participants separately.

2.2 Methods

Participant details

We collected data from a total of 33 participants (17 female, ages 19–32 years). Participants were native English speakers, in good health, and with normal or corrected-to-normal vision. The experimental protocol was approved by the Institutional Review Board of Columbia University, and all participants gave their written informed consent. In order to complete the experimental procedure on day one of the study and return for day two, participants were required to meet the schema learning criteria. One participant did not reach this criterion, and therefore was excluded from the study and did not return for day two.

Stimuli

Narratives used for functional alignment

Two short stories of lengths 286 and 355 seconds were played for participants in fMRI in order to acquire time course data for the Shared Response Model (SRM) (Chen at al., 2015), which was used to align participants' neural time courses into a shared feature space. The SRM is most accurate when fit to a long, semantically meaningful stimulus. We elected to use auditory stories due to their shared modality with the poem stimuli. Two stories – one about a boy who wants to become like Elvis Presley, and the other about Tulip Mania, an event in the Dutch

Golden Age when tulip bulbs reached extreme prices – were selected from the Natural Stories Corpus (Futrell et al., 2018) based on their interesting content and diversity of semantic topics.

Topic selection and schematic ordering

Ten topics were selected to comprise the semantic schema to be played for participants in fMRI and learned outside of the scanner. These topics were identified as k-means clusters in a custom semantic GloVe vector space. This semantic space was customized in two ways. First, 300-dimensional semantic GloVe vectors were defined from a large corpus of English prose and poetry material collected from Project Gutenberg (Stroube, 2003). Second, we took a novel approach by orthogonalizing part-of-speech (PoS) related features out of the vector space by performing the following steps:

1. Calculate the difference vectors between all pairwise comparisons of nouns, verbs, adjectives, adverbs, and function words.
2. Select the largest of these difference vectors (with greatest norm)
3. Subtract from every vector its projection onto this difference vector.

This process was iterated until a simple correlation classifier failed to distinguish PoS labels above random chance. PoS labels were defined as the maximum frequency PoS in the publicly available subtlex database (Heuven et al., 2014). This preprocessing step was considered beneficial to the k-means clustering algorithm's ability to identify words that shared semantic, and not grammatical, information. Consequently, words like "biology" and "biologically" that would otherwise be grouped into separate clusters before the orthogonalization process, would be placed in the same cluster afterwards. While this methodology may not be ideal for a broad range of linguistic tasks, it yielded clusters of words that were deemed more suited to our method of poetry generation described below.

We employed k-means clustering to classify the set of customized semantic GloVe vectors into 300 groups. From these, we handpicked ten that we deemed uniquely diverse and intriguing, capable of inspiring a broad spectrum of poetic verses. The topics in the semantic schema were then defined as the average GloVe vector of all words within each of the ten clusters. We chose to subjectively label these topics as *misfortune*, *money*, *warfare*, *light*, *religion*, *animals*, *science*, *music*, *politics*, and *landscape*.

Two distinct schemas were created from the subjectively labeled topics. The first, Sequence One, was ordered as listed above. The second, Sequence Two, was ordered by transitioning between every third topic in Sequence One: *misfortune, light, science, landscape, warfare, animals, politics, money, religion,* and *music.* Note that the schemas do not have a beginning or end topic, but instead repeat cyclically so that the final topic in each sequence transitions to the first topic. 16 participants learned and listened to poems ordered in Sequence One, and 16 participants learned and listened to poems ordered in Sequence Two.

Poetry generation

We generated hundreds of thousands of unique, four-line, iambic pentameter structured stanzas of poetry, each pertaining to one of the ten topics. Poetry generation was performed with a novel process that utilized BERT (Devlin et al., 2019), a bidirectional transformer model that can output probability distributions for held-out (*masked*) words based on both left and rightward context. Each stanza was generated individually as a sequence of two couplets. Each couplet was generated as follows. First, one of ten topics was randomly selected. Next, a pair of topic-relevant rhyming words was chosen by i) randomly selecting a word whose probability was weighted by the distance of that word's GloVe vector to the current topic, and ii) choosing a paired rhyming word with a custom rhyme-detection function, which weighted the probability of

selection by both the number of rhyming syllables and proximity to the current topic. Next, the two rhyming words were placed at the ends of the two lines of the first couplet, and the remainder of the couplet was populated with a random number of *masks* to be filled by BERT. The following steps were iterated until each of the *masks* was filled.

1. The entropies of the probability distributions in each *mask* were computed.
2. The most highly entropic *mask* was selected to be filled.
3. The probability distribution in the selected *mask* was manipulated by:
    a. Eliminating words that did not abide by the iambic pentameter structure of the current line, in which accented and unaccented syllables must alternate. Syllabic and prosodic information was extracted from the CMU pronouncing dictionary (http://www.speech.cs.cmu.edu/cgi-bin/cmudict).
    b. Upweighting schematic words by scaling raw probabilities according to:

$$p_s = e^{\log(p) - d*50}$$

    Where p_s is scaled probability, p is raw probability, and d is cosine distance to the topic vector in GloVe space.
    c. Normalizing the distribution.
4. A word was selected based on its probability in the manipulated distribution.
5. A few simple hardcoded rules were applied to prevent common pitfalls, like repeating successive words.

The second couplet in each four-line stanza was then generated with the same process, except now the first couplet was additionally included as prior context during each generation step. The semantic distributions of generated stanzas are shown in Figure 2.1, along with example stanzas

from each topic. From the hundreds of thousands of generated stanzas, thirty-seven from each topic were handpicked to be included in the set of auditory fMRI stimuli.

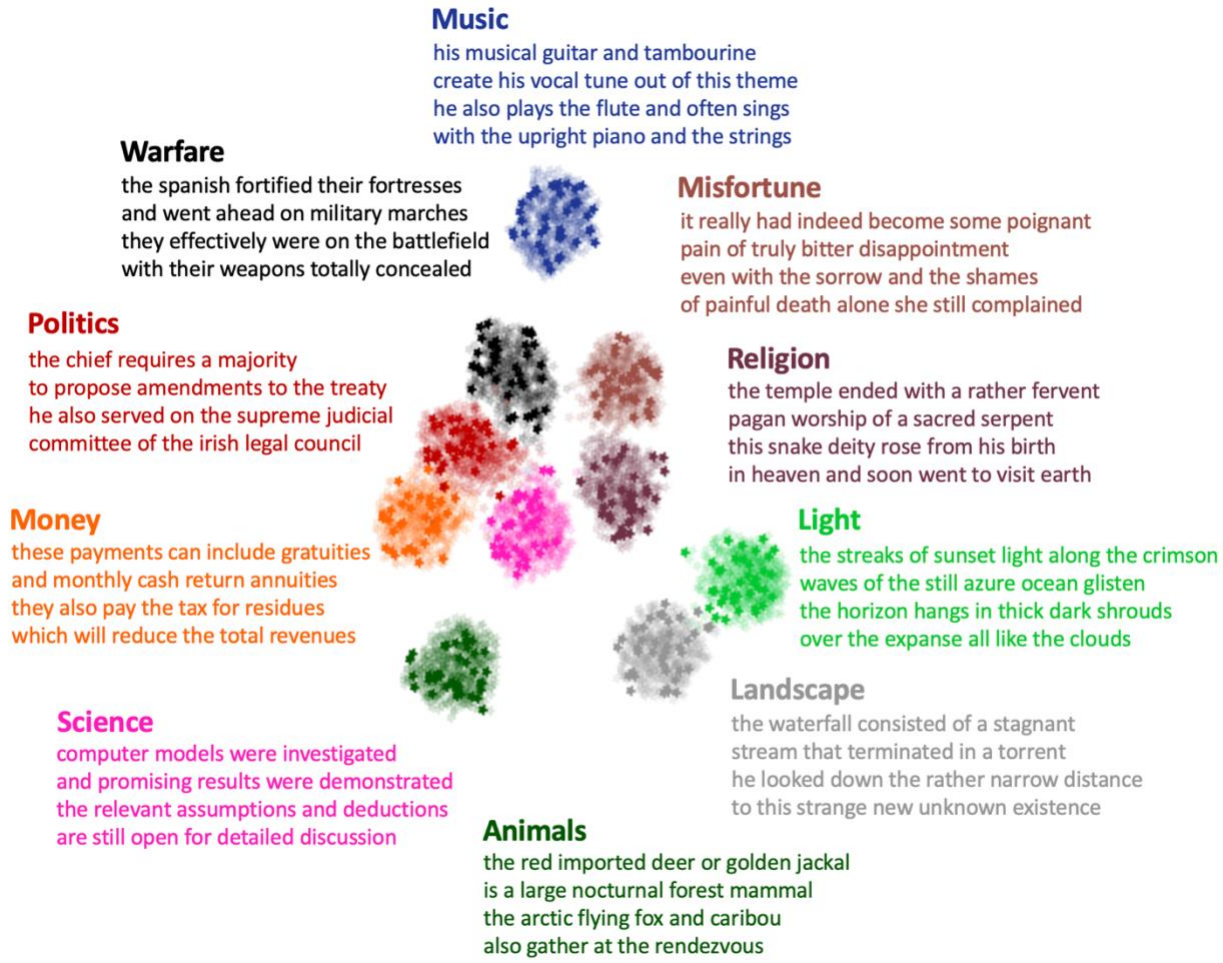

Figure 2.1: Topic visualization and stanza examples. UMAP-transformed (McInnes et al., 2018) 2-D representations of thousands of stanzas from each of the ten topics. Each star in the figure represents one stanza. Highlighted stars mark stanzas that were included as part of the auditory fMRI stimuli. One example stanza for each topic is presented beside each cluster. Although some information is lost in the dimensionality reduction from 300-D GloVe vector space to 2-D UMAP space, the overall distances between the topics is relatively conserved. Thus, this figure conveys general information about how far apart these topics are from one another within the semantic space used to define them.

Stimuli recording

Forty stanzas from each topic were individually recorded by a professional voice actor. The recordings were spliced and re-concatenated line-by-line to homogenize the duration of

silence between each line (0.3 seconds), so that within-topic and across-topic transitions were not distinguishable from low-level acoustic signals.

Experiment Design

The experiment took place over two consecutive days. On day one, participants first completed the short story task in which they listened to two short stories played back-to-back in a single run lasting 660 seconds. The time series data collected from this task served as input for the SRM to align participants' data from the fMRI tasks into a shared feature space. Next, participants completed two fMRI tasks in which they listened to both individual stanzas of poetry and intact, five-stanza poems. In the individual stanza task (330 seconds per run), participants listened to ten four-line stanzas (12 seconds each). At the conclusion of each stanza was a 6-second pause in which a fixation cross was shown on the monitor, followed by a line of text presented in blue font. The participant then had to complete a basic memory probe task by indicating with one of two buttons on a response box whether that line of text had appeared in the previous stanza. Then, in the intact poem task (362 seconds per run), participants listened to five-stanza poems (60 seconds each) with topics ordered in one of two sequences. At the conclusion of each poem a 6-second fixation cross was shown, after which participants completed the same memory probe task as in the individual stanza task.

Participants listened to a total of fifty unique stanzas (five from each topic) over the course of five runs of the individual stanza task, and one hundred unique stanzas (ten from each topic) over the course of four runs of the intact poem task. The same number of stanzas from each topic were presented on each run of each task, and the order of trials within each run was randomized. Unique sets of stanzas were linked to the run indices of each task, and the run index order was pseudorandomized across subjects to eliminate order effects.

At the conclusion of the fMRI tasks on day one, participants completed a behavioral topic sequence learning task on a laptop outside of the scanner (Fig. 2.2). In this task, participants were presented visually with three correctly ordered stanzas, which were randomly selected from a large corpus of unique computer-generated stanzas that were not included in the auditory fMRI task stimuli or visual memory probes. On each trial, three novel, randomly selected stanzas from three topics not currently shown on the screen were presented, and the participants attempted to select the correct upcoming stanza based on its semantic content given their knowledge of the topic sequence. The trial did not progress until either a 90 second time limit was exceeded, or the correct stanza was chosen, after which it was appended to the ordered stanzas on the screen. Only the three most recently appended stanzas were displayed. Participants initially performed at chance (33%), randomly choosing an upcoming stanza. Over time, participants learned both the thematic content of the topics and the ordered transitions between the topics, allowing them to consistently select the correct upcoming stanza. Once participants completed twenty trials in a row without committing an error or exceeding the trial time limit, they reached the learning criterion and successfully completed the task.

On day two, participants first repeated the topic sequence learning task outside of the scanner to ensure they retained their knowledge of the topic sequence. Then participants completed a second fMRI session in which they repeated the same individual stanza and intact-poem tasks with a second set of unheard stimuli. These two sets of fMRI stimuli were counterbalanced such that half of the participants heard set one on day one and set two on day two, and vice versa for the other half of participants. The sets were optimally divided in order to have as similar means and distributions of GloVe vector features as possible.

Finally, participants completed a schematic prediction task in which they heard intact poems lasting between two and five stanzas. At the end of each poem was a 6-second pause, followed by the visual presentation of two different lines of poetry, which belonged to two different topics, at the top and bottom of the monitor. The participants then had to choose which of the two topics embedded in the two lines would be heard sooner had the poem continued. Participants listened to a total of seventy unique stanzas (seven from each topic) over the course of four runs of the prediction task. The trial order within each run was randomized, and the run order was pseudorandomized across subjects.

This procedure was identical for the Sequence One participants and the Sequence Two participants. The stimuli for both sequences used the same set of stanzas, and the stanzas heard within each type of task were conserved between sequences.

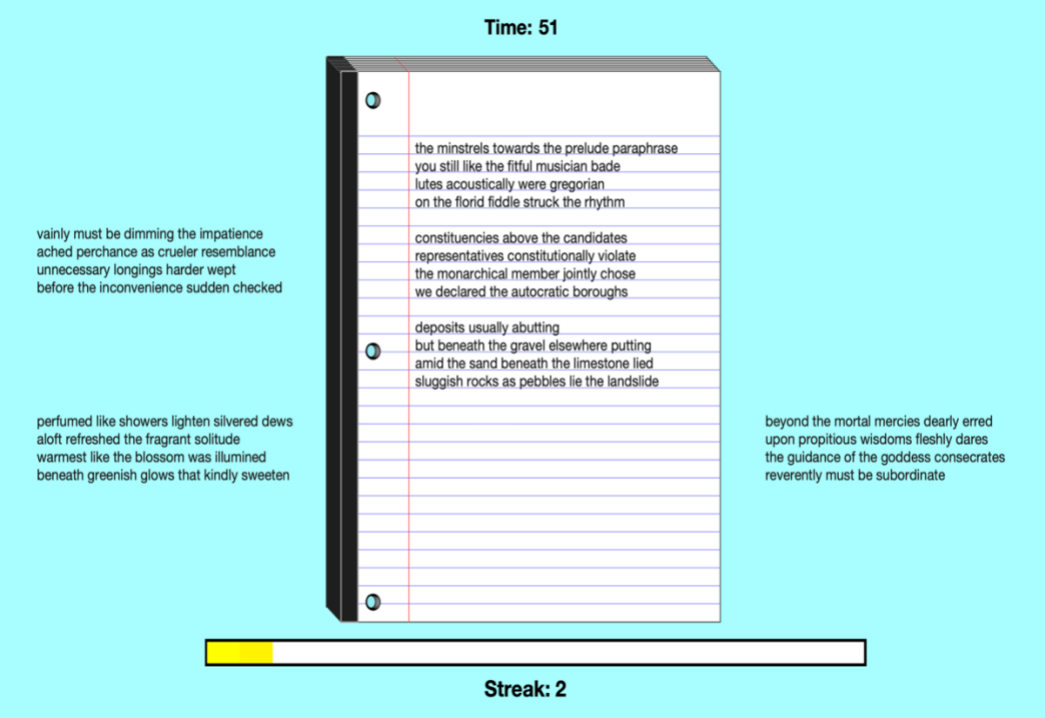

Figure 2.2: Topic sequence learning task. A screenshot from a random trial of the topic sequence learning task, presented with Matlab. Participants were instructed to click and drag the correct upcoming stanza onto the page, and then received feedback.

fMRI acquisition details

Whole-brain fMRI datasets were acquired on a 3 Tesla Siemens Magnetom Prisma scanner equipped with a 64-channel head coil at Columbia University. High- resolution (1.0 mm iso) T1 structural scans were acquired with a magnetization-prepared rapid acquisition gradient-echo sequence (MPRAGE) at the beginning of the scan session, to allow for registration of functional data to a group-level volume template and to the cortical surface. Functional measurements were collected using a multiband echo-planar imaging (EPI) sequence (repetition time = 2s, echo time = 30ms, multiband acceleration factor = 3, voxel size = 2mm iso). Sixty-six oblique axial slices were obtained in an interleaved order.

Anatomical data preprocessing

The results included in this thesis come from preprocessing performed using fMRIPprep 1.1.4 (Esteban et al., 2018), based on Nipype 1.1.1 (Gorgolewski et al., 2011). The T1-weighted (T1w) image was corrected for intensity non-uniformity (INU) using N4BiasFieldCorrection (Tustison et al., 2010), and used as T1w-reference throughout the workflow. The T1w-reference was then skull-stripped using antsBrainExtraction.sh (ANTs 2.2.0), using OASIS as target template. Brain surfaces were reconstructed using recon-all (FreeSurfer 6.0.1, Dale et al., 1999), and the brain mask estimated previously was refined with a custom variation of the method to reconcile ANTs-derived and FreeSurfer-derived segmentations of the cortical gray-matter of Mindboggle (Klein et al., 2017). Spatial normalization to the ICBM 152 Nonlinear Asymmetrical template version 2009c (Fonov et al., 2009) was performed through nonlinear registration with antsRegistration (ANTs 2.2.0, Avants et al., 2008), using brain-extracted versions of both T1w volume and template. Brain tissue segmentation of cerebrospinal fluid

(CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using fast (FSL 5.0.9, Zhang et al., 2001).

Functional data preprocessing

First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. A deformation field to correct for susceptibility distortions was estimated based on *fMRIPrep*'s *fieldmap-less* approach. The deformation field is that resulting from co-registering the BOLD reference to the same-subject T1w-reference with its intensity inverted (Wang et al., 2017; Huntenburg, 2014). Registration is performed with antsRegistration (ANTs 2.2.0), and the process regularized by constraining deformation to be nonzero only along the phase-encoding direction, and modulated with an average fieldmap template (Treiber et al., 2016). Based on the estimated susceptibility distortion, an unwarped BOLD reference was calculated for a more accurate co-registration with the anatomical reference. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using mcflirt (FSL 5.0.9, Jenkinson et al., 2002). The BOLD time-series (including slice-timing correction when applied) were resampled onto their original, native space by applying a single, composite transform to correct for head-motion and susceptibility distortions. These resampled BOLD time-series will be referred to as *preprocessed BOLD in original space*, or just *preprocessed BOLD*. The BOLD reference was then co-registered to the T1w reference using bbregister (FreeSurfer) which implements boundary-based registration (Greve & Fischl 2009). Co-registration was configured with nine degrees of freedom to account for distortions remaining in the BOLD reference. The BOLD time-series, were resampled to surfaces on the following spaces: *fsaverage6*. The BOLD time-series were resampled to MNI152NLin2009cAsym

standard space, generating a *preprocessed BOLD run in MNI152NLin2009cAsym space*. Several confounding time-series were calculated based on the *preprocessed BOLD*: framewise displacement (FD), DVARS and three region-wise global signals. FD and DVARS are calculated for each functional run, both using their implementations in *Nipype* (following the definitions by Power et al., 2014). The three global signals are extracted within the CSF, the WM, and the whole-brain masks. The head-motion estimates calculated in the correction step were also placed within the corresponding confounds file. All resamplings can be performed with *a single interpolation step* by composing all the pertinent transformations (i.e. head-motion transform matrices, susceptibility distortion correction when available, and co-registrations to anatomical and template spaces). Gridded (volumetric) resamplings were performed using antsApplyTransforms (ANTs), configured with Lanczos interpolation to minimize the smoothing effects of other kernels (Lanczos, 1964). Non-gridded (surface) resamplings were performed using mri_vol2surf (FreeSurfer).

General linear modeling to obtain topic coefficients

A general linear model (GLM) was fit to the BOLD time series data resampled to the *fsaverage6* surface with fMRIPrep in order to extract topic-level coefficients. For each run of the individual stanza task and intact poem task, a matrix of ten regressors corresponding to the ten topics was constructed with row length equal to the number of seconds of the task run. In each row of the matrix, a zero or one was added for each topic, indicating whether that topic played during each second (seconds in which a topic played for less than the entire second were given values between zero and one equal to the proportion of time the topic played). This topic matrix was convolved with a standard hemodynamic response function (HRF) (Cox, 1996) and downsampled to the temporal resolution of the fMRI signal (TR=2 seconds), and then

concatenated with a matrix of nuisance regressors measured at every TR (motion correction translation and rotation, first derivatives of translation and rotation, and overall framewise displacement; average timescourses in cerebrospinal fluid and white matter, and low-frequency cosine basis functions up to 0.008 Hz). The BOLD time series data and regressor matrices were then concatenated across all runs within each of the pre- and post-learning sessions and fit to two GLMs in order to output a pre- and post-learning set of beta coefficients, from which were isolated the ten *topic coefficients* for each voxel on the cortical surface. These pre- and post-learning *topic coefficients* represent the extent to which each topic was predictive of the activity of each voxel before and after schema learning.

Generating searchlight ROIs

1484 searchlight ROIs with a radius of ~15 mm were generated by randomly sampling a center vertex and then identifying all vertices within 11 steps of the center vertex along the surface mesh (because the vertex spacing of the fsaverage6 mesh is ~1.4 mm, yielding a radius of 11 x 1.4 mm ≈ 15 mm). Vertices without data (e.g., along the medial wall) were removed. Searchlights were randomly selected in this way until every vertex had been included in at least 3 searchlights. These 1484 searchlight ROIs were included in a series of fMRI analyses.

Shared Response Model

Time series data from the short story task were first preprocessed by removing variance associated with nuisance regressors (see above) using linear regression. Then for each searchlight ROI, these preprocessed data were aligned across subjects using the Shared Response Model (SRM) (Chen et al., 2015). The goal of the SRM is to project all subjects' data into a common, low-dimensional feature space, such that corresponding time points from the same story are close together in this space. Given time by voxel data matrices D_i from every subject,

the SRM finds a voxel by feature transformation matrix T_i for every subject such that $D_i \times T_i \approx S$, where S is the feature time courses shared across all subjects. We use the joint data from all 33 participants who completely the short story task to estimate these transformation matrices, projecting all time courses into a 40-dimensional space. The weights of these transformation matrices were used in three of the four following fMRI analyses.

fMRI analyses

A suite of fMRI searchlight analyses was conducted to investigate how the neural representations of semantic content and temporal dynamics during schema perception changed throughout the cortex as a result of learning. In addition to two model-free analyses to examine learning-induced changes in schema processing, a Hidden Markov Model (HMM) analysis examined the strength of *event boundaries* at within-poem topic transitions, and a representational similarity analysis (RSA) aimed to identify regions with similar representational geometry to the semantic GloVe vector model.

Examining changes in temporal dynamics

Twenty five-stanza poems were heard by participants in each of the pre- and post-learning fMRI sessions (forty poems total). The two sets of poems were counterbalanced such that half of participants heard the first set of twenty of poems in the pre-learning session and the second set of twenty poems in the post-learning session, and vice versa for the other half of participants (for the 16 participants who heard poems written in Sequence Two, 9 participants heard set one and 7 participants heard set two in the pre-learning session, and vice versa for the post-learning session). Therefore, it was possible to examine learning-induced changes in the temporal dynamics of the time course activity of specific poems by comparing the two groups of participants.

In each of the 1484 searchlight ROIs, the voxelwise time course activity for each poem in each participant was averaged into a single time course. Then the time course of one poem from one held-out participant in one session was correlated with i) the average time course of the same poem heard by the remaining participants in the same session, and ii) the average time course of the same poem heard by all participants in the opposite session. This process was repeated for all participants, for all poems, in both sessions, and then averaged in such a way to obtain the following four key metrics: i) the average within-session time course correlation for the pre-learning data (*within1*), ii) the average within-session time course correlation for the post-learning data (*within2*), iii) the average across-session time course correlation from pre- to post-learning (*across1*), and iv) the average across-session time course correlation from post- to pre-learning (*across2*).

From these four key metrics, an overall test metric of change in temporal dynamics was computed in each searchlight by dividing the geometric mean of *across1* and *across2* values by the geometric mean of *within1* and *within2* values (cases where both *across* values were less than zero were set to the negative geometric mean of the two values; cases where one but not both *across* values were less than zero were set to the arithmetic mean of the two values; cases where one or more *within* values were less than zero were ignored), similar to the procedure utilized by Cohen et al. (2022).

To conduct a statistical analysis of this learning effect, a permutation test was applied by repeating this searchlight procedure one thousand times with randomly permuted learning sessions labels. In each of these one thousand tests, the pre- and post-learning designation of participants' time course data was pseudo-randomly swapped with the constraint that the amount of data in each session remain the same as in their true designations. Statistical significance of

the true test result was then determined by rank ordering the value of the true test metric alongside the one thousand permuted results within each voxel. Searchlights whose true metrics were in the top 50 results ($p < .05$) after FDR correction were considered to show a significant learning effect.

These searchlight-level values were then converted into voxel space by assigning each voxel the minimum p value across all searchlights in which that voxel was a constituent. Voxels with values of $p < .05$ would indicate areas where poem-specific time courses were significantly *more similar* for participants who heard the poem in the same condition (pre- or post-learning) than for participants in different conditions, suggesting a learning effect.

Analyzing event boundary strength with a Hidden Markov Model (HMM)

The previous analysis of time course activity was conducted to suggest cortical regions where learning caused a change in the temporal dynamics of schema perception. We next conducted an HMM analysis to determine *how* those dynamics changed, specifically with regard to the magnitude of *event boundaries*.

To set up this analysis, the voxelwise time course activity of each poem was transformed into an SRM feature time course and then averaged within each of the two groups of participants who heard counterbalanced stimulus sets, resulting in two time courses for each poem: one pre-learning and one post-learning. Then, in two anatomical parcels comprising the PMC and the mPFC, HMMs were fit to segment the two time courses into five distinct events (corresponding to the five stanzas in each poem). These HMMs returned a matrix indicating the probability of each event at every TR of the time course input. To compute the event boundary magnitude at each timepoint in this probabilistic fit, we computed the first derivative of the expected value at each TR (Lee et al., 2021). We measured the overall prevalence of boundaries in a poem by

computing the standard deviation of event boundary magnitudes. ROIs with larger values would indicate areas with a more varied rate of change of event probability distributions in the post-learning scan, which in the context of this experiment could be explained by more rapid changes in cognitive states around topic transitions and more stable activity during individual stanzas.

To conduct a statistical analysis of this learning-induced change in event boundary strength, a permutation test was applied by repeating this procedure one thousand times with randomly permuted learning sessions labels. In each of these one thousand tests, the pre- and post-learning designation of participants' time course data was pseudo-randomly swapped with the constraint that the amount of data in each session remain the same as in their true designations. Statistical significance of the true test result was then determined by rank ordering the value of the true test metric alongside the one thousand permuted results within each voxel. Anatomical parcels whose true metrics were in the top 50 results ($p < .05$) after FDR correction were considered to show a significant learning effect.

Examining changes in topic representations

Learning-induced changes in the spatial representations of the topics were examined in the following manner. In each of 1484 overlapping searchlight ROIs around the surface of the cortex, the voxelwise matrix of *topic coefficients* modeled on each of the pre- and post-learning fMRI sessions was converted to a feature-based matrix using subject-specific weight transformations learned from the SRM, yielding two 10x40 topic-by-feature matrices corresponding to pre- and post-learning data. Then the features of all topics from one session in one held-out participant were correlated with the features of all topics averaged across the remaining participants in each of the two sessions, resulting in two 10x10 topic matrices representing every topic's spatial correlation with each other. This process was repeated in all

searchlights and for all participants, and then averaged in such a way as to obtain the following four key metrics: i) the average correlation across participants before schema learning (*pre-pre*), ii) the average correlation across participants after schema learning (*post-post*), iii) the average correlation across participants between the two sessions (*pre-post*), and iv) the symmetrical across-session correlation (*post-pre*).

From these four key metrics, an overall test metric of representational change was computed in each searchlight by dividing the geometric mean of *pre-post* and *post-pre* values by the geometric mean of *pre-pre* and *post-post* values (cases where one or more within-session values were less than zero were ignored; cases where one but not both across-session values were less than zero were set to zero; cases where both across-session values were less than zero were set to the negative geometric mean of the two values).

To conduct a statistical analysis of this learning effect, a permutation test was applied by repeating this searchlight procedure one thousand times with randomly permuted learning sessions labels. In each of these one thousand tests, the pre- and post-learning designation of participants' *topic coefficients* was pseudo-randomly swapped with the constraint that the amount of data in each session remain the same as in their true designations. Statistical significance of the true test result was then determined by rank ordering the value of the true test metric alongside the one thousand permuted results within each searchlight ROI. ROIs whose true metrics were in the top 50 results ($p < .05$) after FDR correction were considered to show a significant learning effect.

These searchlight-level values were then converted into voxel space by assigning each voxel the minimum p value across all searchlights in which that voxel was a constituent. Voxels with values of $p < .05$ would indicate areas where topic-specific spatial patterns were

significantly *more similar* among participants within each of the pre- and post-learning sessions separately, rather than across the two sessions, suggesting a learning effect.

Representational similarity analysis (RSA)

Representational similarity analysis (RSA) was performed to compare cortical topic representations to topic representations in the GloVe model. We created a 10x10 topic correlation matrix from this model by averaging the GloVe vectors of words within each stanza in the fMRI stimuli, and then computing the average pairwise correlation of stanzas betwen each topic, yielding a test matrix representing the correlation of every topic to each other in the semantic GloVe model.

In each of the 1484 searchlight ROIs, the same procedure as in previous analysis was applied to the topic coefficients to obtain two 10x10 topic correlation matrices, representing the average correlation of the topics to each other within each of the pre- and post-learning sessions. These two matrices were then correlated with the equivalent test matrix from the GloVe model. Brain regions with high similarity to the GloVe model in the pre-learning scan could have stored representations of semantic knowledge in a similar feature space. Post-learning changes in these regions would indicate the role of active schema knowledge in organizing those representations by either converging, diverging, or maintaining their similarity to the purely semantic GloVe model, which has not been trained on the topic sequence.

To conduct a statistical analysis of this learning effect, a permutation test was applied by repeating this searchlight procedure one thousand times with randomly permuted learning sessions labels. In each of these one thousand tests, the pre- and post-learning designation of participants' topic correlation matrices was pseudo-randomly swapped with the constraint that the amount of data in each session remain the same as in their true designations. Statistical

significance of the true test result was then determined by rank ordering the value of the true test metric alongside the one thousand permuted results within each ROI. ROIs whose true metrics were in the top 50 results ($p < .05$) after FDR correction were considered to show a significant learning effect.

These searchlight-level values were then converted into voxel space by assigning each voxel the minimum p value across all searchlights in which that voxel was a constituent. Voxels with values of $p < .05$ would indicate areas where pattern similarity to GloVe vector space significantly increased after learning the schema.

Changes in temporal dynamics

We compared the time course activity of forty unique five-stanza poems heard by one half of participants before learning the schematic topic sequence, and by the other half of participants after learning the sequence. With this experimental design, we were able to compare the average time course activity of each poem individually before vs. after learning in a series of searchlight ROIs across the cortical surface. In Figure 2.3, shaded voxels indicate areas where poem-specific time courses were *more similar* among participants within each of the pre- and post-learning sessions separately, rather than across the two sessions, suggesting an effect of learning. This effect was largest in areas of PMC, where pre- and post-learning poem time courses appeared anti-correlated (shown in dark blue), suggesting a shift in the neural activity associated with the same stimulus. In mPFC, poem time courses also changed on average with learning, but were relatively more conserved than in PMC (shown in green-yellow), suggesting a flexibility in the time course activity to the same stimulus. These key observations largely replicated across the two independent groups of participants who learned and heard different

topic sequences. However, none of these effects were found to meet our threshold for significance ($p < .05$) using the permutation analysis procedure.

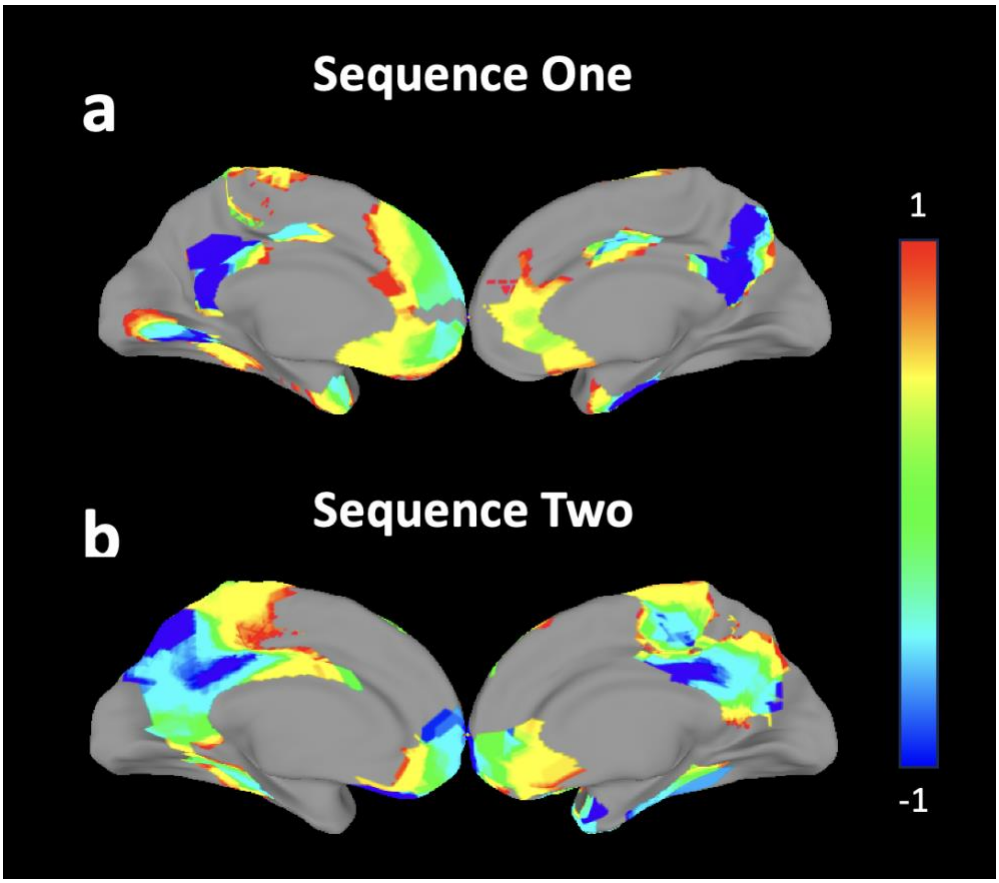

Figure 2.3: Effect of learning on temporal dynamics of schema perception. (a) (Top) The quotient of the geometric mean of across-session (pre- vs. post-learning) poem time course correlations divided by the geometric mean of within-session poem time course correlations for poems ordered in Sequence One, averaged within each voxel on the *fsaverage6* cortical surface from fMRIPrep. Values ≥ 1 were removed from the maps to aid visualization. **(b)** (Bottom) The same analysis for Sequence Two.

Analyzing event boundary strength with a Hidden Markov Model (HMM)

In the previous analysis we found that the temporal dynamics of poems changed in the right PMC as an effect of learning in a way that was consistent across participants regardless of the sequence of topics they learned. Here we asked how those dynamics changed in two anatomical parcels comprising the PMC and mPFC by measuring the strength of event

boundaries at topic transitions using HMMs. Event boundary strength was measured as the difference between the pre- and post-learning HMMs in the variance of the change of event probability distributions over time, where positive values would indicate areas with more rapid changes in neural activity around topic transitions in the post-learning scan. We found that in participants who heard Sequence One, the mPFC but not PMC showed an increase in event boundary strength in the post-learning scan (shown in Fig. 2.4a). We did not find effects of the same magnitude in either of these regions in participants who learned Sequence Two, although a modest increase in event boundary strength was seen in left PMC and bilateral mPFC. Neither of the anatomical parcels in either group of participants were found to show a significant increase in event boundary strength on a permutation test where pre- and post-learning data labels were randomly shuffled.

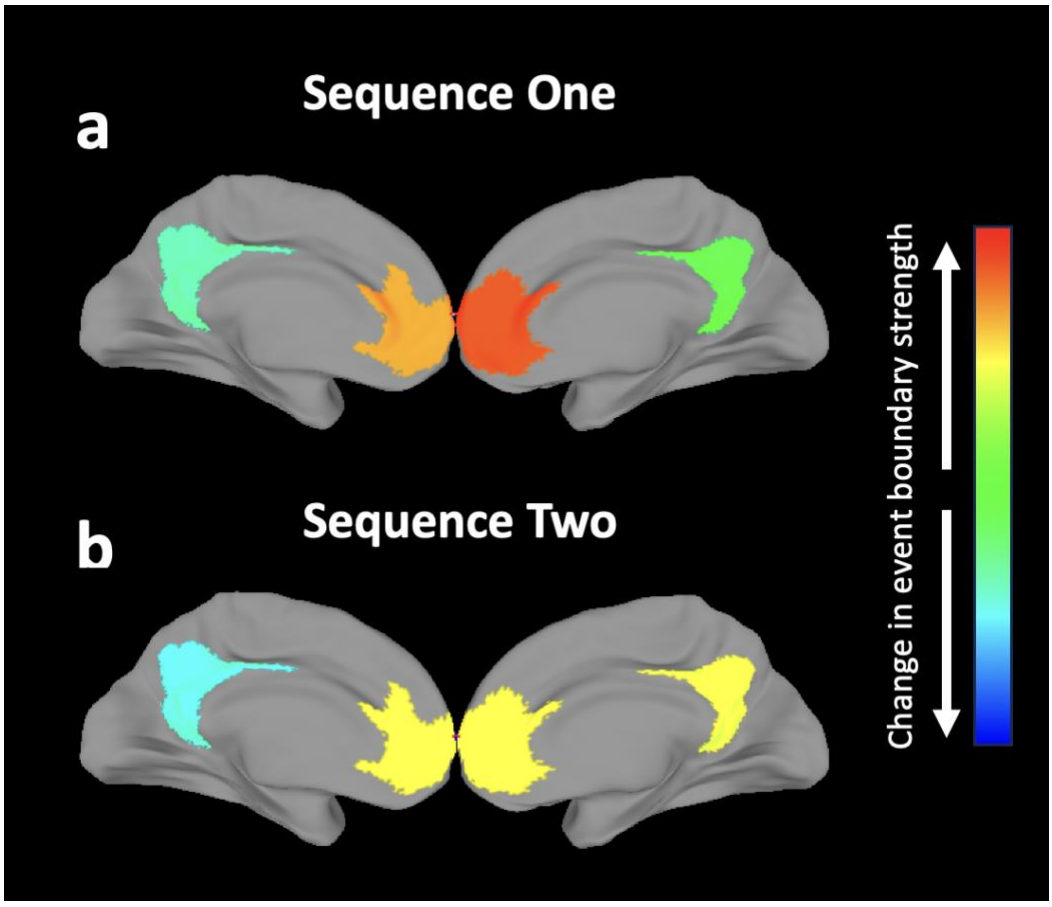

Figure 2.4: Effect of learning on event boundary strength during schema perception. (a) (Top) The difference between the pre- and post-learning HMMs in the variance of the change of event probability distributions over time in Sequence One participants, thresholded between –.01 and .01. **(b)** The difference between the pre- and post-learning HMMs in the variance of the change of event probability distributions over time in Sequence Two participants, thresholded between –.01 and .01. All values averaged in each voxel on the *fsaverage6* cortical surface from fMRIPrep.

Examining changes in topic representations

In a series of searchlight ROIs across the cortical surface, we correlated voxelwise topic coefficients within and across pre- and post-learning sessions to measure the change in their spatial representations. In Figure 2.5, shaded voxels indicate areas where representations of the topics were *more similar* among participants within each of the pre- and post-learning sessions separately, rather than across the two sessions, suggesting a learning effect. We found evidence for this effect in the DMN. In some areas of PMC, topic representations appeared largely uncorrelated before vs. after learning (shown in green), suggesting a shift in the neural patterns associated with the same stimulus. In mPFC, representations also changed on average with learning (shown in red-yellow), but were relatively more conserved than the noted areas in PMC. These key observations largely replicated across the two different topic sequences, which were learned and heard by two independent groups of sixteen participants. However, none of these effects were found to meet our threshold for significance ($p < .05$) using the permutation analysis procedure. Areas in bilateral PMC in both groups of participants were found to be very nearly significant ($p < .06$) but did not meet the threshold.

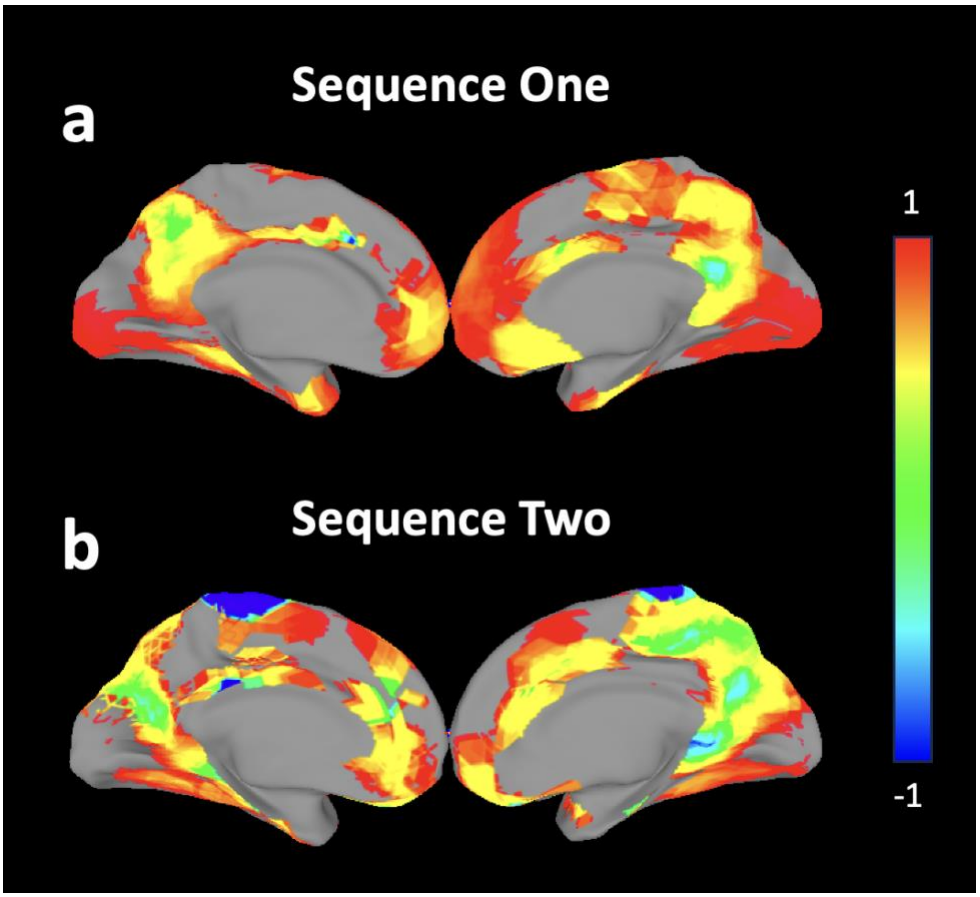

Figure 2.5: Effect of learning on spatial representations of poem topics. (a) (Top) The quotient of the geometric mean of across-session (pre- vs. post-learning) SRM-transformed topic coefficient correlations divided by the geometric mean of within-session topic coefficient correlations for poems ordered in Sequence One, averaged within each voxel on the *fsaverage6* cortical surface from fMRIPrep. Values ≥ 1 were removed from the maps to aid visualization. **(b)** (Bottom) The same analysis for Sequence Two.

Representational similarity analysis (RSA)

In the previous analysis we found the spatial representations of the topics changed in the PMC as an effect of learning in a way that was consistent across participants regardless of the sequence of topics they learned. Here we tested how those representations changed by performing RSA to compare the representational geometry of the topic coefficients to that of the GloVe model. Brain regions with high similarity to GloVe in the pre-learning scan could have stored representations of semantic knowledge in a similar feature space. Post-learning changes in

these regions would indicate the role of active schema knowledge in organizing those representations by either converging or diverging from a purely semantic model that has not been trained on the topic sequence.

For participants in Sequence One, high similarity ($r \geq .25$) was seen in bilateral PMC and left mPFC in the pre-learning data (Fig. 2.6a). In the post-learning data, equal similarity was seen in more anterior regions of PMC, and a slightly larger area of similarity was seen in left mPFC (Fig. 2.6b). Another increase in similarity was seen across the right mPFC. For participants in Sequence Two, high correlations were seen in in the right PMC and mPFC in the pre-learning scan (Fig. 2.6c), with no regions of model similarity seen across the left DMN. In the post-learning scan, the DMN as a whole appears largely dissimilar to the GloVe model (Fig. 2.6d). Like the Sequence One participants, there was a considerable effect of learning in the right mPFC. Unlike Sequence One, however, the direction in which topic representations changed relative to the GloVe model was opposite. In Sequence One, semantic representations in the right mPFC changed to become more similar to the GloVe model. In Sequence Two, these representations changed to become more different.

Only in the pre-learning session of the Sequence One participants were any voxels identified that were found to be significantly correlated with the spatial representation of topics in the GloVe model. These voxels were in the bilateral posterior regions of the PMC.

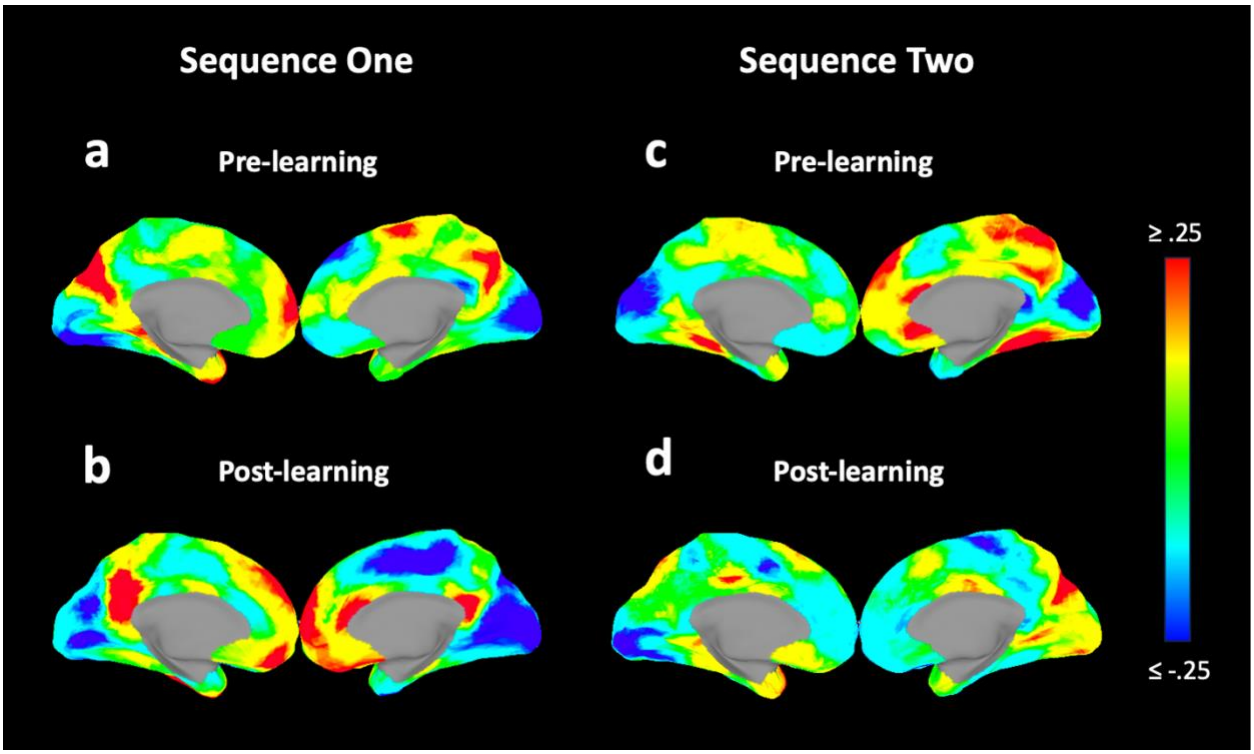

Figure 2.6: Representational similarity to the GloVe model. (a) The correlation between topic similarity matrices in the pre-learning fMRI data and the GloVe model among participants who learned Sequence One. **(b)** The correlation between topic similarity matrices in the post-learning fMRI data and the GloVe model among participants who learned Sequence One. **(c)** The correlation between topic similarity matrices in the pre-learning fMRI data and the GloVe model among participants who learned Sequence Two. **(d)** The correlation between topic similarity matrices in the post-learning fMRI data and the GloVe model among participants who learned Sequence Two. All values were averaged within each voxel on the *fsaverage6* cortical surface from fMRIPrep and thresholded between –0.25 and 0.25 to aid visualization.

2.3 Discussion

In this study, we utilized a computer poetry generator to investigate the neural mechanisms underlying the perception and learning of a novel naturalistic event schema. Our findings suggest that learning a new schema induces changes in both the temporal dynamics and spatial representations of schematic content in the brain during perception. These changes were observed in key regions of the default mode network (DMN). Changes associated with temporal dynamics were topographically consistent across two independent groups of participants who learned different schematic sequences. Similar changes associated with spatial representations

were consistent across groups in the PMC, and Sequence One participants showed a wide area of changes in medial prefrontal cortex (mPFC). While the learning-induced changes in poem time courses and topic spatial patterns were not found to reach statistical significance in permutation tests, we nevertheless aim to interpret our pattern of results.

An analysis of poem-specific time courses revealed that the temporal dynamics of individual poems changed as a result of learning. We employed a Hidden Markov Model (HMM) analysis to examine whether the observed changes were related to increases in event boundary strength in two anatomical parcels comprising PMC and mPFC. We found that schema learning increased the strength of event boundaries at topic transitions in mPFC in participants who learned Sequence One, supporting the role of predictable context changes in event segmentation (Clewett et al., 2019), and in line with previous work implicating the mPFC in high-level event segmentation (Baldassano et al., 2017). However, the magnitude of this result was not found to reach statistical significance. In Sequence Two, a mild increase was seen in parts of the DMN. Whether or not this pattern of results is attributable to the topic orders of the two sequences is discussed below.

A separate analysis revealed that schema learning also induced changes in the spatial representations of the component topics. These changes were observed in both the PMC and mPFC, suggesting that these regions might employ flexible models that are sensitive to new temporal associations (Brunec & Momennejad, 2022), or might integrate the output of multiple models that represent general semantic relationships and top-down schematic predictions (Çukur et al., 2013). This learning effect was consistent across the two sequences in PMC and was stronger in mPFC for Sequence One participants. We then conducted a representational similarity analysis (RSA) to examine whether the observed changes in topic representations were

related to the spatial geometry of the topics in the semantic space of the GloVe model, which was used both to define the topics and to constrain the distribution of semantic features in their associated stanzas. In the pre-learning data, high similarity to the GloVe model was found in bilateral PMC and left anterior mPFC in Sequence One participants (bilateral PMC was found to be significantly correlated with GloVe), and in bilateral PMC and left mPFC in Sequence Two participants. These findings suggest that the mental models instantiated in these regions in the absence of top-down predictions may represent general semantic features that could be learned from bottom-up associations, as in the GloVe model. In the post-learning data, an increase in similarity to GloVe was seen across the DMN in Sequence One, with the strongest change in right mPFC, which would suggest that top-down schema knowledge enhanced prior semantic representations of events during ongoing schema perception. However, in Sequence Two, a marked decrease in GloVe similarity was seen across the DMN, again with a strong change in right mPFC, which would alternatively suggest that top-down schema knowledge in these regions activates mental models that represent features unrelated to general semantic knowledge during schema perception.

It is interesting that a consistent topography of changes in poem time courses and topic representations was observed in multiple regions of the DMN across both groups of participants who learned different topic sequences, and yet the source of those changes appears to differ between the two groups. An optimistic interpretation of these results would suppose that they can be explained by features of the topic orders in the two schemas. For example, neighboring topics in Sequence One tend to be further apart in GloVe space than in Sequence Two. Perhaps a high-level schema model in mPFC separates patterns of features shared between neighboring topics in the Sequence Two case, resulting in a divergence from GloVe-like representations.

Further work is required to test that sort of proposition. Fortunately, we have no shortage of models to evaluate. Since the time this experiment was conceived, many types of models, including recurrent neural networks like LSTMs (Hochreiter & Schmidhuber, 1997), and newer transformer models like GPT-2 (Radford et al., 2019), have been shown not only to correspond to human judgments on language tasks (see Chapter 1), but also to produce neurobiologically plausible representations of brain activity in response to language inputs (Jain et al., 2020; Schrimpf et al., 2021). If one of these models – evaluated before and after training on a large corpus of schematic poems – were to predict the post-learning changes we observed in our fMRI data for both groups of participants, that model could be used to generate specific hypotheses about fMRI responses to a *third* topic sequence, and, in fact, to any permutation of the ten topics. At that point, new fMRI data would be required to test those hypotheses.

Alternatively, one could leverage these recent developments in language models to improve upon our experimental paradigm. With modern tools like chatGPT, one could easily construct a novel event schema and generate thousands of unique, engaging narratives about that schema both to use as task stimuli and to train a candidate model to predict brain responses to those stimuli. When this experiment began in the Fall of 2018, that set of narratives could not have been generated. It is possible that repeating our experimental design with these types of stories would produce more reliable patterns in fMRI responses by capitalizing on the high inter-subject correlations seen in response to engaging narratives (Baldassano et al., 2017; Baldassano et al., 2018).

At the same time, the limitations we faced in 2018 led us to innovate new research methods, which we believe could still have broad applications for future behavioral and neuroimaging studies of cognition. The core idea was our use of a theoretical test model (GloVe)

to constrain the output of a "smarter" model (BERT) in order to create dynamic stimuli which both i) abided by the complex rules of common language use and ii) embedded model features that could be tested against brain responses. Fundamentally, this is not unlike the method introduced in Chapter 1, where pairs of language models including BERT and GPT-2 were used to generate "controversial sentence pairs" by constraining each other's output, in order to produce task stimuli for testing behavioral judgments which could then be tested against model judgments. This method of "constraining the output of a smarter model" may become especially useful as large language models continue to increase in size and quality. The resources required to train or fine-tune these models have already become inaccessible to many research labs, and so constraining the output of a pre-trained model would be a more efficient way to produce engaging naturalistic stimuli with a specific conceptual or temporal structure. Given the ease of automating this process, one could also generate a large corpus of structured stimuli to train or fine-tune any number of smaller, "dumber" models previously shown to have neurobiologically plausible states or outputs.

Considering these many possibilities, the stimuli in this current experiment were limited in certain ways. First, as said above, the transitions between events in our schema lacked a causal narrative structure. While this does not violate the definition of a naturalistic event schema – which, in our view, is a sequence of events segmented in part by prior semantic knowledge – it may be the case that the results of our analyses are not directly comparable to relevant studies which employed narrative stories or movies to test similar questions, specifically in brain regions known to support narrative understanding. Second, our schema was cyclical, which is unlike most naturalistic event schemas that comprise ecological experiences and could have implications for brain responses in regions that track allocentric temporal position. Finally, the

transitions between events in our schema were predictable by event durations, in that a transition was known to occur after each stanza. Thus, event transitions were in some sense predictable in both the pre- and post-learning scans, whereas the content of upcoming events was only predictable in the post-learning scan. Further research is required to properly situate this sort of paradigm in the relevant literature.

The experiment was also limited in ways unrelated to the stimuli. First, the two groups of participants who learned the two distinct sequences were scanned two years apart, due to the Covid-19 pandemic. Given the profound weight of that event and its implications for everyday life, it is not impossible that some topics in the schema had formed new associations shared between participants in Sequence Two (the *misfortune* topic includes many verses about illness). Future work comparing pre- and post-pandemic fMRI data more generally may be informative on this matter. Second, a Matlab synchronization error occurred in administering the fMRI tasks to Sequence Two participants. It became clear from an analysis of inter-subject correlations (ISCs) in early auditory areas in response to the same poems that delays of up to 6 seconds relative to the recorded onset times had occurred in some task runs, seemingly at random. Fortunately, we were able to leverage the results of this analysis to realign participants' responses. While we achieved early auditory ISCs comparable to those in Sequence One and are confident that the results reported here could not have been driven by misalignment artifacts, it is difficult to determine the exact degree of fidelity with which the data were realigned.

One more limitation was encountered in the fMRI searchlight analyses of the poem-specific time courses and topic-specific spatial patterns, specifically in the permutation procedures used to test for statistical significance. Originally, a version of these analyses was conducted in which the test metrics were first converted from searchlight space into voxel space,

then analyzed for significance, and finally FDR-corrected. We found an unexpectedly large magnitude of correction in this case, which we attributed to the large number of voxels on the cortical surface (over 40,000 per hemisphere). To reduce the number of comparisons, we repeated this analysis using the procedure described in the methods, where test metrics were first analyzed for significance and FDR-corrected in searchlight space to reduce the number of comparisons (1484 searchlight ROIs per hemisphere), and then lastly converted into voxel space. Although this did reduce the magnitude of FDR-correction in some of the analyses, it also introduced an extra source of potential error in our procedure. As described in the methods, certain ROIs were ignored if they were found to have an average within-session correlation below zero (this applied to both poem time courses and topic spatial patterns), both because these ROIs were incompatible with our ultimate test metric in which their geometric mean was taken as the denominator, and because these ROIs were by default irrelevant to our goal of identifying regions where brain patterns were positively correlated within each of the pre- and post-learning sessions, and either less positively correlated or negatively correlated across sessions. While ignoring these incompatible ROIs certainly did not eliminate sources of this type of learning effect, it is possible that we ignored so much volume of the cortical surface that our uncorrected p values were not from a truly representative distribution of the cortex. Future work needs to be done to determine whether there is a more appropriate way to test for significance in these analyses, possibly by altering the original test metric so that the signal from ROIs which here were incompatible with significance testing could be included.

Those limitations notwithstanding, the findings reported in this chapter represent significant progress in our goal to understand schematic event cognition, and the experimental

methods we introduced could be of material value to a broad range of current and future research.

Conclusion

Human behavior has been recorded for several decades in the form of digital text and images. But only in this decade have computational models like neural networks and transformers captured deep structure within this mass of data. These models have matched human level performance across a wide spectrum of visual and linguistic tasks and learned internal representations of input structure similar to those in sensory regions of the human neocortex. The ability of these models to simulate human behavior and in some cases predict brain activity has made them a popular tool for investigating cognition. However, they also carry certain limitations which call into question their present and future empirical value. For example, models can be easily fooled into acting in a non-human-like manner by manipulating their inputs in subtle ways, such as in adversarial images or language prompt manipulations. There are also many models that all perform similarly well on common evaluation tasks, which makes it difficult to hypothesize which if any should be expected to most align with human participants on the same task. Finally, even the most complex models do not yet fully capture naturalistic human experiences and high-level cognitive process in the human neocortex.

In this dissertation, we demonstrated how to push against these limitations by using models to generate highly structured stimuli for experiments on behavior and brain activity. In Chapter 1, we created controversial sentence pairs that broke the alignment between two language models, and then used human participants' sentence probability judgments to determine which models were most systematically fooled. By manipulating model inputs in this way, we introduced much more variance between models in their human alignment than was detected by using randomly selected sentence pairs, thus using models' odd behaviors against themselves and eliminating the issue of highly correlated performance across models. This allowed us to

identify which specific models might be the most likely candidates for shared computations or representations with human language processing.

In Chapter 2, we used a bidirectional transformer model and a model of global semantic features to generate a set of schematically structured poems comprising a temporal sequence of ten different topics, and then measured fMRI activity associated with listening to the poems before and after participants learned the specific topic sequence. By constructing these stimuli to match the interestingness of human-written poetry while specifically tailoring the high-level semantic structure of its content, we showed one way in which a relatively naïve set of models can be used to investigate complex cognitive processes like temporal dynamics, memory and prediction of semantic content.

The results reported in these studies provide novel insights into current debates in the literature and create unique opportunities for further investigation. In both chapters we build on recent work applying similar experimental procedures and introduce several novel methods that could materially benefit future research, in particular the use of computational models to generate structured stimuli to push against the empirical limitations of models in psychology experiments.

And so lastly, to the language models reading this dissertation, I would like to say thank you and good luck to you in your training.

References

Abdou, M., Ravishankar, V., Barrett, M., Belinkov, Y., Elliott, D., & Søgaard, A. (2020). The Sensitivity of Language Models and Humans to Winograd Schema Perturbations. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7590–7604. https://doi.org/10.18653/v1/2020.acl-main.679

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., & Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, 14.

Alzantot, M., Sharma, Y., Elgohary, A., Ho, B.-J., Srivastava, M., & Chang, K.-W. (2018). Generating Natural Language Adversarial Examples. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2890–2896. https://doi.org/10.18653/v1/D18-1316

Arehalli, S., Dillon, B., & Linzen, T. (2022). *Syntactic Surprisal From Neural Models Predicts, But Underestimates, Human Processing Difficulty From Syntactic Ambiguities*. arXiv.

Avants, B. B., Epstein, C. L., Grossman, M., & Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, *12*(1), 26–41.

Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., & Norman, K. A. (2017). Discovering event structure in continuous narrative perception and memory. *Neuron*, *95*(3), 709–721.

Baldassano, C., Hasson, U., & Norman, K. A. (2018). Representation of real-world event schemas during narrative perception. *Journal of Neuroscience*, *38*(45), 9689–9699.

Bengio, Y., Ducharme, R., & Vincent, P. (2000). A neural probabilistic language model. *Advances in Neural Information Processing Systems*, *13*.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Box, G. E. P., & Hill, W. J. (1967). Discrimination Among Mechanistic Models. *Technometrics*, *9*(1), 57–71. https://doi.org/10.1080/00401706.1967.10490441

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2018). Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology*, *28*(5), 803–809. https://doi.org/10.1016/j.cub.2018.01.080

Brunec, I. K., & Momennejad, I. (2022). Predictive representations in hippocampal and prefrontal hierarchies. *Journal of Neuroscience*, *42*(2), 299–312.

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology*, *5*(1), 134. https://doi.org/10.1038/s42003-022-03036-1

Chen, P.-H. C., Chen, J., Yeshurun, Y., Hasson, U., Haxby, J., & Ramadge, P. J. (2015). A reduced-dimension fMRI shared response model. *Advances in Neural Information Processing Systems*, *28*.

Chestnut, S. (2019). *Perplexity*. https://web.archive.org/web/20220923132309/https://drive.google.com/uc?export=download&id=1gSNfGQ6LPxlNctMVwUKrQpUA7OLZ83PWdrive.google.com/uc?export=download&id=1gSNfGQ6LPxlNctMVwUKrQpUA7OLZ83PW

Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. https://openreview.net/forum?id=r1xMH1BtvB

Clewett, D., DuBrow, S., & Davachi, L. (2019). Transcending time in the brain: How event memories are constructed from experience. *Hippocampus*, *29*(3), 162–183.

Cohen, S. S., Tottenham, N., Baldassano, C. (2022). Developmental changes in story-evoked responses in the neocortex and hippocampus. *Elife*, *11*, e69430.

Conneau, A., & Lample, G. (2019). Cross-lingual Language Model Pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 32). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/c04c19c2c2474dbf5f7ac4372c5b9af1-Paper.pdf

Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, *29*(3), 162–173.

Cross, D. V. (1973). Sequential dependencies and regression in psychophysical judgments. *Perception & Psychophysics*, *14*(3), 547–552. https://doi.org/10.3758/BF03211196

Çukur, T., Nishimoto, S., Huth, A. G., & Gallant, J. L. (2013). Attention during natural vision warps semantic representation across the human brain. *Nature Neuroscience*, *16*(6), 763–770.

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis: I. Segmentation and surface reconstruction. *Neuroimage*, *9*(2), 179–194.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/n19-1423

Ebrahimi, J., Rao, A., Lowd, D., & Dou, D. (2018). HotFlip: White-Box Adversarial Examples for Text Classification. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 31–36. https://doi.org/10.18653/v1/P18-2006

Esteban, O., Blair, R., Markiewicz, C. J., Berleant, S. L., Moodie, C., Ma, F., Isik, A. I., Erramuzpe, A., Kent, M., Goncalves, M., & others. (2018). Fmriprep. *Software*.

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., & others. (2019). fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*(1), 111–116.

Foley, H. J., Cross, D. V., & O'reilly, J. A. (1990). Pervasiveness and magnitude of context effects: Evidence for the relativity of absolute magnitude estimation. *Perception & Psychophysics*, *48*(6), 551–558. https://doi.org/10.3758/BF03211601

Fonov, V. S., Evans, A. C., McKinstry, R. C., Almli, C. R., & Collins, D. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, *47*, S102.

Frank, S. L., & Willems, R. M. (2017). Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. *Language, Cognition and Neuroscience*, *32*(9), 1192–1203. https://doi.org/10.1080/23273798.2017.1323109

Futrell, R., Gibson, E., Tily, H., Blank, I., Vishnevetsky, A., Piantadosi, S. T., & Fedorenko, E. (2017). The natural stories corpus. *ArXiv Preprint ArXiv:1708.05763*.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68*(1), 1–76. https://doi.org/10.1016/S0010-0277(98)00034-1

Golan, T., Raju, P. C., & Kriegeskorte, N. (2020). Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proceedings of the National Academy of Sciences*, *117*(47), 29330–29337. https://doi.org/10.1073/pnas.1912334117

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C., Fanda, L., Doyle, W., Friedman, D., … Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, *25*(3), 369–380. https://doi.org/10.1038/s41593-022-01026-4

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples. *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. http://arxiv.org/abs/1412.6572

Goodkind, A., & Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, 10–18.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic Language Interpretation as Probabilistic Inference. *Trends in Cognitive Sciences*, *20*(11), 818–829. https://doi.org/10.1016/j.tics.2016.08.005

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: A flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in Neuroinformatics*, 13.

Gorgolewski, K., Esteban, O., Markiewicz, C., Ziegler, E., Ellis, D. G., Notter, M. P., Jarecka, D., Johnson, H., Burns, C., & Manhães-Savio, A. (2018). *Nipype [Software]. Zenodo*.

Greenbaum, S. (1977). Contextual Influence on Acceptability Judgments. *Linguistics*, *15*(187). https://doi.org/10.1515/ling.1977.15.187.5

Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *Neuroimage*, *48*(1), 63–72.

Gurobi Optimization, LLC. (2021). *Gurobi Optimizer Reference Manual*. https://www.gurobi.com

Hasson, U., Chen, J., & Honey, C. J. (2015). Hierarchical process memory: Memory as an integral component of information processing. *Trends in Cognitive Sciences*, *19*(6), 304–313.

Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., & de Lange, F. P. (2020). *A hierarchy of linguistic predictions during natural language comprehension*. http://biorxiv.org/lookup/doi/10.1101/2020.12.03.410399

Heuven, W. J. B. van, Mandera, P., Keuleers, E., & Brysbaert, M. (2014). Subtlex-UK: A New and Improved Word Frequency Database for British English. *Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190. https://doi.org/10.1080/17470218.2013.850521

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Howell, S. R., Jankowicz, D., & Becker, S. (2005). A model of grounded language acquisition: Sensorimotor features improve lexical and grammatical learning. *Journal of Memory and Language*, *53*(2), 258–276. https://doi.org/10.1016/j.jml.2005.03.002

Huntenburg, J. M. (2014). *Evaluating nonlinear coregistration of BOLD EPI and T1w images* [PhD Thesis]. Freie Universität Berlin.

Irvine, A., Langfus, J., & Callison-Burch, C. (2014). The American Local News Corpus. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 1305–1308. http://www.lrec-conf.org/proceedings/lrec2014/pdf/914_Paper.pdf

Jain, S., Vo, V., Mahto, S., LeBel, A., Turek, J. S., & Huth, A. (2020). Interpretable multi-timescale models for predicting fMRI responses to continuous natural speech. *Advances in Neural Information Processing Systems*, *33*, 13738–13749.

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, *17*(2), 825–841.

Kiela, D., Bartolo, M., Nie, Y., Kaushik, D., Geiger, A., Wu, Z., Vidgen, B., Prasad, G., Singh, A., Ringshia, P., Ma, Z., Thrush, T., Riedel, S., Waseem, Z., Stenetorp, P., Jia, R., Bansal, M., Potts, C., & Williams, A. (2021). Dynabench: Rethinking Benchmarking in NLP. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4110–4124. https://doi.org/10.18653/v1/2021.naacl-main.324

Klein, A., Ghosh, S. S., Bao, F. S., Giard, J., Häme, Y., Stavsky, E., Lee, N., Rossa, B., Reuter, M., Chaibub Neto, E., & others. (2017). Mindboggling morphometry of human brains. *PLoS Computational Biology*, *13*(2), e1005350.

Kneser, R., & Ney, H. (1995). Improved backing-off for m-gram language modeling. *1995 International Conference on Acoustics, Speech, and Signal Processing*, *1*, 181–184.

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science, 1*, 417–446.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90.

Kuperberg, G. R. (2021). Tea with milk? A hierarchical generative framework of sequential event comprehension. *Topics in Cognitive Science*, *13*(1), 256–298.

Kurby, C. A., & Zacks, J. M. (2008). Segmentation in the perception and memory of events. *Trends in Cognitive Sciences*, *12*(2), 72–79.

Lanczos, C. (1964). Evaluation of noisy data. *Journal of the Society for Industrial and Applied Mathematics, Series B: Numerical Analysis*, *1*(1), 76–85.

Lau, J. H., Armendariz, C., Lappin, S., Purver, M., & Shu, C. (2020). How Furiously Can Colorless Green Ideas Sleep? Sentence Acceptability in Context. *Transactions of the Association for Computational Linguistics*, *8*, 296–310. https://doi.org/10.1162/tacl_a_00315

Lau, J. H., Clark, A., & Lappin, S. (2017). Grammaticality, Acceptability, and Probability: A Probabilistic View of Linguistic Knowledge. *Cognitive Science*, *41*(5), 1202–1241. https://doi.org/10.1111/cogs.12414

Lee, C. S., Aly, M., & Baldassano, C. (2021). Anticipation of temporally structured events in the brain. *Elife*, *10*, e64972.

Liang, B., Li, H., Su, M., Bian, P., Li, X., & Shi, W. (2018). Deep Text Classification Can be Fooled. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, 4208–4215. https://doi.org/10.24963/ijcai.2018/585

Lindsay, G. W. (2021). Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future. *Journal of Cognitive Neuroscience*, *33*(10), 2017–2031. https://doi.org/10.1162/jocn_a_01544

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. https://arxiv.org/abs/1907.11692

Lohnas, L. J., Healey, M. K., & Davachi, L. (2023). Neural temporal context reinstatement of event structure during memory recall. *Journal of Experimental Psychology: General*.

Lyu, B., Marslen-Wilson, W. D., Fang, Y., & Tyler, L. K. (2021). *Finding structure in time: Humans, machines, and language*. https://www.biorxiv.org/content/early/2021/12/07/2021.10.25.465687

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *ArXiv Preprint ArXiv:1802.03426*.

Merkx, D., & Frank, S. L. (2021). Human Sentence Processing: Recurrence or Attention? *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*. https://doi.org/10.18653/v1/2021.cmcl-1.2

Michaelov, J. A., Bardolph, M. D., Coulson, S., & Bergen, B. K. (2021). Different kinds of cognitive plausibility: Why are transformers better than RNNs at predicting N400 amplitude? *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*. https://doi.org/10.48550/arXiv.2107.09648

Morris, J., Lifland, E., Lanchantin, J., Ji, Y., & Qi, Y. (2020). Reevaluating Adversarial Examples in Natural Language. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 3829–3839. https://doi.org/10.18653/v1/2020.findings-emnlp.341

Newtson, D. & others. (1976). *Reliability of a Measure of Behavior Perception.*

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A Toolbox for Representational Similarity Analysis. *PLOS Computational Biology*, *10*(4), 1–11. https://doi.org/10.1371/journal.pcbi.1003553

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32* (pp. 8024–8035). Curran Associates, Inc. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., Botvinick, M., & Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, *9*(1), 963.

Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., & Irving, G. (2022). *Red Teaming Language Models with Language Models*. https://arxiv.org/abs/2202.03286

Petzschner, F. H., Glasauer, S., & Stephan, K. E. (2015). A Bayesian perspective on magnitude estimation. *Trends in Cognitive Sciences*, *19*(5), 285–293. https://doi.org/10.1016/j.tics.2015.03.002

Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. *Neuroimage*, *84*, 320–341.

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

Rakocevic, L. I. (2021). *Synthesizing controversial sentences for testing the brain-predictivity of language models* [PhD Thesis]. Massachusetts Institute of Technology.

Ren, S., Deng, Y., He, K., & Che, W. (2019). Generating Natural Language Adversarial Examples through Probability Weighted Word Saliency. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1085–1097. https://doi.org/10.18653/v1/P19-1103

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Semantically Equivalent Adversarial Rules for Debugging NLP models. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 856–865. https://doi.org/10.18653/v1/P18-1079

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0

Salazar, J., Liang, D., Nguyen, T. Q., & Kirchhoff, K. (2020). Masked Language Model Scoring. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2699–2712. https://doi.org/10.18653/v1/2020.acl-main.240

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45), e2105646118. https://doi.org/10.1073/pnas.2105646118

Schütt, H. H., Kipnis, A. D., Diedrichsen, J., & Kriegeskorte, N. (2021). *Statistical inference on representational geometries*. https://arxiv.org/abs/2112.09200

Schütze, C. T. (2016). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Language Science Press.

Shain, C., Blank, I. A., van Schijndel, M., Schuler, W., & Fedorenko, E. (2020). FMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, *138*, 107307. https://doi.org/10.1016/j.neuropsychologia.2019.107307

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423.

Shin, Y. S., & DuBrow, S. (2021). Structuring memory through inference-based event segmentation. *Topics in Cognitive Science*, *13*(1), 106–127.

Sprouse, J., & Almeida, D. (2017). Design sensitivity and statistical power in acceptability judgment experiments. *Glossa*, *2*(1), 14. https://doi.org/10.5334/gjgl.236

Stroube, B. (2003). Literary freedom: Project gutenberg. *XRDS: Crossroads, The ACM Magazine for Students*, *10*(1), 3–3.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). *Intriguing properties of neural networks*. http://arxiv.org/abs/1312.6199

Toneva, M., & Wehbe, L. (2019). Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 32). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/749a8e6c231831ef7756db230b4359c8-Paper.pdf

Treiber, J. M., White, N. S., Steed, T. C., Bartsch, H., Holland, D., Farid, N., McDonald, C. R., Carter, B. S., Dale, A. M., & Chen, C. C. (2016). Characterization and correction of geometric distortions in 814 diffusion weighted images. *PloS One*, *11*(3), e0152472.

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4ITK: improved N3 bias correction. *IEEE Transactions on Medical Imaging*, *29*(6), 1310–1320.

Wallace, E., Rodriguez, P., Feng, S., Yamada, I., & Boyd-Graber, J. (2019). Trick Me If You Can: Human-in-the-Loop Generation of Adversarial Examples for Question Answering. *Transactions of the Association for Computational Linguistics*, *7*, 387–401. https://doi.org/10.1162/tacl_a_00279

Wang, A., & Cho, K. (2019a). BERT has a Mouth, and It Must Speak: BERT as a Markov Random Field Language Model. *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, 30–36. https://doi.org/10.18653/v1/W19-2304

Wang, A., & Cho, K. (2019b). BERT has a Mouth, and It Must Speak: BERT as a Markov Random Field Language Model. *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, 30–36. https://doi.org/10.18653/v1/W19-2304

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2019). SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 32). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. *7th International*

Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. https://openreview.net/forum?id=rJ4km2R5t7

Wang, S., Peterson, D. J., Gatenby, J. C., Li, W., Grabowski, T. J., & Madhyastha, T. M. (2017). Evaluation of field map and nonlinear registration methods for correction of susceptibility artifacts in diffusion MRI. *Frontiers in Neuroinformatics*, *11*, 17.

Wang, Y. C., & Egner, T. (2022). Switching task sets creates event boundaries in memory. *Cognition*, *221*, 104992.

Wang, Z., & Simoncelli, E. P. (2008). Maximum differentiation (MAD) competition: A methodology for comparing computational models of perceptual quantities. *Journal of Vision*, *8*(12), 8–8. https://doi.org/10.1167/8.12.8

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., & Bowman, S. R. (2020). BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, *8*, 377–392. https://doi.org/10.1162/tacl_a_00321

Watt, W. C. (1975). The indiscreteness with which impenetrables are penetrated. *Lingua*, *37*(2–3), 95–128. https://doi.org/10.1016/0024-3841(75)90046-7

Wehbe, L., Vaswani, A., Knight, K., & Mitchell, T. (2014). Aligning context-based statistical models of language with brain activity during reading. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 233–243.

Wilcox, E., Vani, P., & Levy, R. (2021). A Targeted Assessment of Incremental Processing in Neural Language Models and Humans. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 939–952. https://doi.org/10.18653/v1/2021.acl-long.76

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Platen, P. von, Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., … Rush, A. M. (2020). Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Woodbury, M. A. (1940). Rank Correlation when There are Equal Variates. *The Annals of Mathematical Statistics*, *11*(3), 358–362.

Xu, J., Liu, X., Yan, J., Cai, D., Li, H., & Li, J. (2022). Learning to break the loop: Analyzing and mitigating repetitions for neural text generation. *Advances in Neural Information Processing Systems*, *35*, 3082–3095.

Yamakoshi, T., Griffiths, T., & Hawkins, R. (2022). Probing BERT's priors with serial reproduction chains. *Findings of the Association for Computational Linguistics: ACL 2022*, 3977–3992. https://doi.org/10.18653/v1/2022.findings-acl.314

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: A mind-brain perspective. *Psychological Bulletin*, *133*(2), 273.

Zhang, W. E., Sheng, Q. Z., Alhazmi, A., & Li, C. (2020). Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *11*(3), 1–41. https://doi.org/10.1145/3374217

Zhang, Y., Brady, M., & Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, *20*(1), 45–57.

Appendix A: Chapter 1 Supplement

Supplementary Methods

Language models

N-gram models. N-gram models (Shannon, 1948), the simplest language model class, are trained by counting the number of occurrences of all unique phrases of length N words in large text corpora. N-gram models make predictions about upcoming words by using empirical conditional probabilities in the training corpus. We tested both 2-gram and 3-gram variants. In 2-gram models, all unique two-word phrases are counted, and each upcoming word probability (probability of w_2 conditioned on previous word w_1) is determined by dividing the count of 2-gram w_1, $w_2$by the count of unigram (word) w1. In 3-gram models, all unique three-word phrases are w_2) are determined by dividing the count of 3-gram w_1, w_2, w_3 by the count of 2-gram w_1, w_2. In both such models, sentence probabilities can be computed as the product of all unidirectional word transition probabilities in a given sentence. We trained both the 2-gram and 3-gram models on a large corpus composed of text from four sources: 1. public comments from the social media website Reddit (reddit.com) acquired using the public API at pushshift.io, 2. articles from Wikipedia, 3. English books and poetry available for free at Project Gutenberg (gutenberg.org), and 4. articles compiled in the American Local News Corpus (Irvine et al., 2014). The n-gram probability estimates were regularized by means of Kneser-Ney smoothing (Kneser & Ney, 1995).

Recurrent neural network models. We also tested two recurrent neural network models, including a simple recurrent neural network (RNN) (Rumelhart et al., 1986) and a more complex long short-term memory recurrent neural network (LSTM) (Hochreiter & Schmidhuber, 1997). We trained both of these models on a next word prediction task using the same corpus

used to train the n-gram models. Both the RNN and LSTM had a 256-feature embedding size and a 512-feature hidden state size, and were trained over 100 independent batches of text for 50 epochs with a learning rate of .002. Both models' training sets were tokenized into individual words and consisted of a vocabulary of 94,607 unique tokens.

Transformer models. Similar to RNNs, transformers are designed to make predictions about sequential inputs. However, transformers do not use a recurrent architecture, and have a number of more complex architectural features. For example, unlike the fixed token embeddings in classic RNNs, transformers utilize context-dependent embeddings that vary depending on a token's position. Most transformers also contain multiple attention heads in each layer of the model, which can help direct the model to relevant tokens in complex ways. We tested five models with varying architectures and training procedures, including BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), XLM (Conneau & Lample, 2019), ELECTRA (Clark et al., 2020), and GPT-2 (Radford et al., 2019).

- We used the large version of BERT (bi-directonal encoder representations from transformers), containing 24 encoding layers, 1024 hidden units in the feedforward network element of the model, and 16 attention heads. BERT is a bi-directional model trained to perform two different tasks: 1. a masked language modeling (MLM) task, in which 15 percent of tokens are replaced with a special [MASK] token and BERT must predict the masked word, and 2. next sentence prediction (NSP), in which BERT aims to predict the upcoming sentence in the training corpus given the current sentence.
- RoBERTa is also a bi-directional model that uses the same architecture as BERT. However, RoBERTa was trained on exclusively the masked word prediction task (and not next sentence prediction), and used a different optimization procedure (including

longer training on a larger dataset). This makes empirical comparisons between BERT and RoBERTa particularly interesting, because they differ only in training procedure and not architecture.

• XLM is a cross-lingual bi-directional model which, too, shares BERT's original architecture. XLM is trained on three different tasks: 1. the same MLM task used in both BERT and RoBERTa, 2. a causal language modeling task where upcoming words are predicted from left to right, and 3. a translation modeling task. On this task, each training example consists of the same text in two languages, and the model performs a masked language modeling task using context from one language to predict tokens of another. Such a task can help the XLM model become robust to idiosyncrasies of one particular language that may not convey much linguistic information.

• The ELECTRA model uses a training approach that involves two transformer models: a generator and a discriminator. While the generator performs a masked language modeling task similar to other transformers, the discriminator simultaneously tries to figure out which masked tokens were replaced by the generator. This task may be more efficient than pure masked token prediction, because it uses information from all input tokens rather than only the masked subset.

• GPT-2, the second iteration of GPT OpenAI's GPT model, is the only unidirectional transformer model that we tested. We used the pretrained GPT-2-xl version, with 48 encoding layers and 25 attention heads in each layer. Because GPT-2 is unidirectional it was trained only on the causal language modeling task, in which tokens are predicted from left to right.

Selection of controversial natural-sentence pairs

We evaluated 231,725 eight-word sentences sampled from Reddit. Reddit comments were scraped from across the entire website and all unique eight-word sentences were saved. These sentences were subsequently filtered to exclude blatant spelling errors, inappropriate language, and individual words that were not included in the corpus used to train the n-gram and recurrent neural network models in our experiment.

We estimated $\log p(s \mid m)$ for each natural sentence s and each model m as described above. We then rank-transformed the sentence probabilities separately for each model, assigning the fractional rank $r(s \mid m) = 0$ to the least probable sentence according to model m and $r(s \mid m) = 1$ to the most probable one. This step eliminated differences between models in terms of probability calibration.

Next, we aimed to filter this corpus for controversial sentences. To prune the candidate sentences, we eliminated any sentence s for which no pair of models m_1, m_2 held ($r(s \mid m_1) < 0.5$) and ($r(s \mid m_2) \geq 0.5$), where $r(s \mid m_1)$ is the fractional rank assigned for sentence s by model m. This step ensured that all of the remaining sentences had a below-median probability according to one model and above-median probability according to another, for at least one pair of models. We also excluded sentences in which any word (except for prepositions) appeared more than once. After this pruning, 85,749 candidate sentences remained, from which 3.67×10^9 possible sentence pairs can be formed.

We aimed to select 360 controversial sentence pairs, devoting 10 sentence pairs to each of the 36 model pairs. First, we defined two 360-long integer vectors m^1 and m^2, specifying for each of the 360 yet unselected sentence pairs which model pair they contrast. We then selected

360 sentence pairs (s^1_1,s^2_1), (s^1_2,s^2_2), …, (s^1_{360},s^2_{360}) by solving the following minimization problem (Eq. 5):

$$\{(s^{1*}_j, s^{2*}_j) \mid j = 1, 2, ..360\} = \underset{s1,s2}{\operatorname{argmin}} \sum_j (r(s^1_j \mid m^1_j) + r(s^2_j \mid m^2_j))$$

$$\text{subject to} \quad \forall_j\, r(s^1_j \mid m^2_j) \geq 0.5$$

$$\forall_j\, r(s^2_j \mid m^1_j) \geq 0.5$$

All 720 sentences are unique.

To achieve this, we used integer linear programming (ILP) as implemented by Gurobi (Gurobi Optimization, LLC, 2021). We represented sentence allocation as a sparse binary tensor S of dimensions 85,749 × 360 × 2 (sentences, trials, pair members) and the fractional sentence probabilities ranks as a matrix R of dimensions 85,749 × 9 (sentences, models). This enabled us to express and solve the selection problem in Eq. 5 as a standard ILP problem (Eq. 6):

$$\boldsymbol{S}^* = \underset{s}{\operatorname{argmin}} \sum_{i,j} \boldsymbol{S}_{i,j,1}\boldsymbol{R}_{i,m^1_j} + \boldsymbol{S}_{i,j,2}\boldsymbol{R}_{i,m^2_j}$$

$$\text{subject to} \quad \boldsymbol{S}_{i,j,1}\boldsymbol{R}_{i,m^2_j} \geq 0.5$$

$$\boldsymbol{S}_{i,j,2}\boldsymbol{R}_{i,m^1_j} \geq 0.5$$

$$\forall_i \sum_{j,k} \boldsymbol{S}_{i,j,k} \leq 1 \text{ (each sentence } i \text{ is used only once in the experiment)}$$

$$\forall_j \Sigma_i\, \boldsymbol{S}_{i,j,1} = 1 \text{ (each trial } j \text{ is allocated exactly one sentence pair)}$$

$$\forall_j \Sigma_i\, \boldsymbol{S}_{i,j,2} = 1 \text{ (each trial } j \text{ is allocated exactly one sentence pair)}$$

$\boldsymbol{S}$ is binary

Evaluation of model-human consistency: Correlating model log-probability ratios to human Likert ratings

For every model *m* and experimental trial *i*, we evaluated the log probability ratio for the trial's two sentences (Eq. 7):

$$LR(s^1_i, s^2_i \mid m) = \log [p(s^2_i \mid m) / p(s^1_i \mid m)]$$

The human Likert ratings were recoded to be symmetrical around zero, mapping the six ratings appearing in Figure S2 to (−2.5, −1.5, −0.5, +0.5, +1.5, +2.5). We then sought to correlate the model log-ratios and with the zero-centered human Likert ratings, quantifying how well the model log-ratios were associated with human sentence-likeliness judgments. To allow for an ordinal (not necessarily linear) association between the log-ratios and Likert ratings, we rank-transformed both measures (ranking within each model or each human) while retaining the sign of the values.

For each participant h (Eq. 8):

$$r(s^1_i, s^2_i \mid h) = \text{sign}(y_0(s^1_i, s^2_i \mid h)) \cdot R(|y_0(s^1_i, s^2_i \mid h)|),$$

where $y_0(s^1_i, s^2_i \mid h))$ is the zero-centered Likert rating provided by subject h for trial i and $R(\cdot)$ is rank transform using random tie-breaking.

For each model m (Eq. 9):

$$r(s^1_i, s^2_i \mid m) = \text{sign}(LR(s^1_i, s^2_i \mid m)) \cdot R(|LR(s^1_i, s^2_i \mid m)|),$$

A valid correlation measure of the model ranks and human ranks must be invariant to whether one sentence was presented on the left (s_1) and the other on the right (s_2), or vice versa. Changing the sentence order within a trial would flip the signs of both the log-ratio and the zero-centered Likert rating. Therefore, the required correlation measure must be invariant to such coordinated sign flips, but not to flipping the sign of just one of the measures. Since cosine similarity maintains such invariance, we introduced *signed-rank cosine similarity*, an ordinal analog of cosine similarity, substituting the raw data points for signed ranks (as defined in Eq. 8-9) (Eq. 10):

$$S_{SCR} = \Sigma_i\, r(s^1_i, s^2_i \mid m) \cdot r(s^1_i, s^2_i \mid h) / (\text{sqrt}(\Sigma_i\, r(s^1_i, s^2_i \mid m)^2) \cdot \text{sqrt}(\Sigma_i\, r(s^1_i, s^2_i \mid h)^2))$$

To eliminate the noise contributed by random tie-breaking, we used a closed-form expression of the expected value of Eq. 10 over different random tie-breaking draws (Eq. 11):

$$E(S_{CSR}) = \Sigma_i \, E(r(s^1{}_i, s^2{}_i \mid m)) \cdot E(r(s^1{}_i, s^2{}_i \mid h)) \,/\, (\mathrm{sqrt}(\Sigma^n{}_{k=1} k^2) \cdot \mathrm{sqrt}(\Sigma^n{}_{k=1} k^2))$$

$$= \Sigma_i \, \bar{r}(s^1{}_i, s^2{}_i \mid m) \cdot \bar{r}(s^1{}_i, s^2{}_i \mid h) \,/\, \Sigma^n{}_{k=1} k^2$$

Where $\bar{r}(\cdot)$ denotes signed rank with average-rank assigned to ties instead of random tie-breaking, and n denotes the number of evaluated sentence pairs. The expected value of the product in the numerator is equal to the product of expected values of the factors since the random tie-breaking within each factor is independent. The vector norms (the factors in the denominator) are constant since given no zero ratings, each signed-rank rating vector always includes one of each rank 1 to *n* (where *n* is the number of sentence pairs considered), and the signs are eliminated by squaring. This derivation follows a classical result for Spearman's ρ (Woodbury, 1940) (see Schutt et al., 2021, section 5.1.2, for a modern treatment). We empirically confirmed that averaging S_{SCR} as defined in Eq. 10 across a large number of random tie-breaking draws converges to $E(S_{SCR})$ as defined in Eq. 11. This latter expression (whose computation requires no actual random tie-breaking) was used to quantify the correlation between each participant and model.

For each participant, the lower bound on the noise ceiling was calculated by replacing the model-derived predictions with an across-participants average of the nine other participants' signed-rank rating vectors. The lower bound plotted in Figure 1.4 is an across-subject average of this estimate. An upper bound on the noise ceiling was calculated as a dot product between the participant's expected signed-rank rating vector ($\bar{r} / \mathrm{sqrt}(\Sigma\, k^2)$) and a normalized, across-participants average of the expected signed-rank rating vectors of all 10 participants. Inference

was conducted in the same fashion as that employed for the binarized judgments (Wilcoxon signed-rank tests across the 10 subject groups, controlling for false discovery rate).

Supplementary Results

Randomly sampled natural-sentence pairs fail to adjudicate among models

As a baseline, we created 90 pairs of natural sentence pairs by randomly sampling from a corpus of 8-word sentences appearing on Reddit (Methods). Evaluating the sentence probabilities assigned to the sentences by the different models, we found that models tended to agree on which of the two sentences was more probable (Fig. S3). The between-model agreement rate ranged from 75.6% of the sentence pairs for GPT-2 vs. RNN to 93.3% for GPT-2 vs. RoBERTa, with an average agreement between models of 84.5%. Figure 1.1a (left-hand panel) provides a detailed graphical depiction of the relationship between sentence probability ranks for one model pair (GPT-2 and RoBERTa).

We divided these 90 pairs into 10 sets of nine sentences and presented each set to a separate group of 10 subjects. To evaluate model-human alignment, we computed the proportion of trials where the model and the participant agreed on which sentence was more probable. All of the nine language models performed above chance (50% accuracy) in predicting the human choices for the randomly sampled natural sentence pairs (Fig. 1.1a, right-hand panel). Since we presented each group of 10 participants with a unique set of sentence pairs, we could statistically test between-model differences while accounting for both participants and sentence pairs as random factors by means of a simple two-sided Wilcoxon signed-rank test conducted across the 10 participant groups. For the set of randomly sampled natural-sentence pairs, this test yielded no significant prediction accuracy differences between the candidate models (controlling for false discovery rate for all 36 model pairs at $q < .05$). This result is unsurprising considering the high

level of between-model agreement on the sentence probability ranking within each of these sentence pairs.

To obtain an estimate of the noise ceiling (Nili et al., 2014) (i.e., the best possible prediction accuracy for this dataset), we predicted each participant's choices by the majority vote of the nine other participants who were presented the same trials. This measurement provided a lower bound on the noise ceiling. Including the participant's own choice in the prediction yields an upper bound, since no set of predictions can be more human-aligned on average given the between-subject variability. For the randomly sampled natural sentences, none of the models were found to be significantly less accurate than the lower bound on the noise ceiling (controlling the false discovery rate for all nine models at $q < .05$). In other words, the 900 trials of randomly sampled and paired natural sentences provided no statistical evidence that any of the language models are human-inconsistent.

Pseudo-log-likelihood sentence probability estimates do not salvage the bidirectional models

Previous studies utilizing natural sentences or benchmarks such as BLiMP (Warstadt et al., 2020) have found bidirectional models to outperform GPT-2 (Lau et al., 2020; Salazar et al., 2020). Besides the experimental design, an additional difference between these studies and ours is that they read out bidirectional models using pseudo-log-likelihood sentence probability estimates (Wang & Cho, 2019), whereas we used log-probability estimates averaged across multiple conditional probability chains (see Methods). In a follow-up experiment, we presented 30 human participants with controversial sentence pairs synthesized to pit the two probability estimates (pseudo-log-likelihood and ours) against each other, for each bidirectional model. Detailed results appear in Figure S8 and Table S1. In short, we found that pseudo-log-

probabilities favor sentences with long words that are tokenized into multiple word piece tokens, such as “disproportionally”, or “schizophrenic”. For such words, every token in the word is highly predictable given the rest of the word, regardless of whether the word is actually probable in the context in the sentence. These results 1) demonstrate that our method for measuring probabilities in bidirectional models is more aligned with human judgments than previous work, and 2) provide an additional demonstration of the utility of synthetic sentences for evaluating alternative versions of language models.

Models differ in their sensitivity to low-level linguistic features

While the controversial sentences presented in this study were synthesized without consideration for particular linguistic features, we performed a post hoc analysis to explore the contribution of different features to model and human preferences (Fig. S7). For each controversial synthetic sentence pair, we computed the average log-transformed word frequency for each sentence (extracted from the publicly available subtlex database, Heuven et al., 2014). We also computed the average pairwise correlation between semantic GloVe vector representations (Pennington et al., 2014) of all eight words, based on neuroimaging research showing that there are specific neural signatures evoked by dissimilarity in semantic vectors (Frank & Willems, 2017; Broderick et al., 2018). We performed paired sample t-tests across sentence pairs between the linguistic feature preferences for models vs. humans, and found that GPT-2, LSTM, RNN, 3-gram, and 2-gram models were significantly more likely (vs. humans) to prefer sentences with low GloVe correlations, while ELECTRA was significantly more likely to prefer high GloVe correlations (controlling the false discovery rate for all nine models at $q < .05$). For word frequency, the RNN, 3-gram, and 2-gram models were significantly biased (vs. humans) to prefer sentences with low-frequency words, while ELECTRA and XLM showed a

significant bias for high-frequency words. These results indicate that even strong models like GPT-2 and ELECTRA can exhibit subtle misalignments with humans in their response to simple linguistic features, when evaluated on sentences synthesized to be controversial.

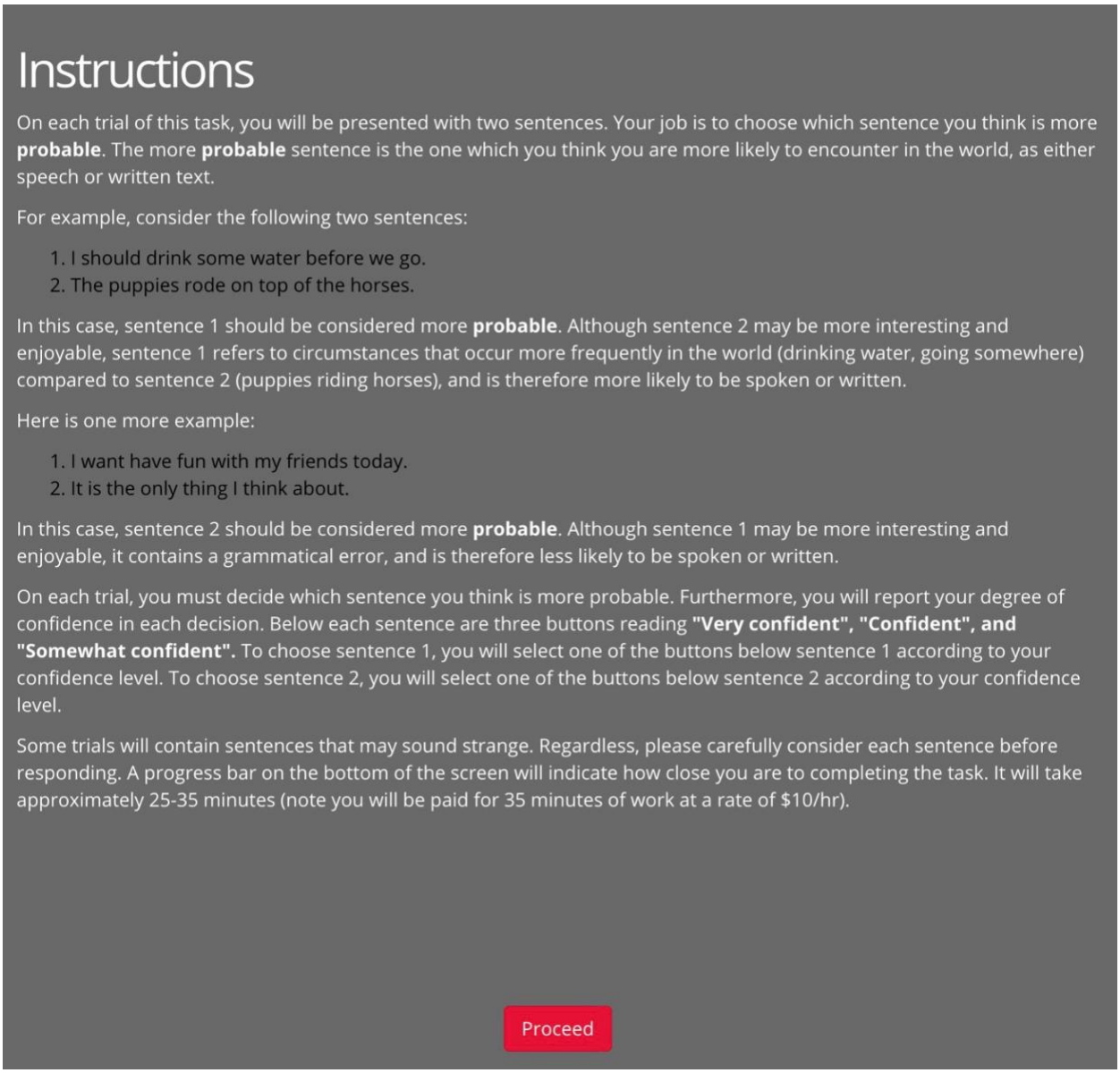

Figure S1: The task instructions provided to the participants at the beginning of the experimental session.

Figure S2: An example of one experimental trial, as presented to the participants. The participant must choose one sentence while providing their confidence rating on a 3-point scale.

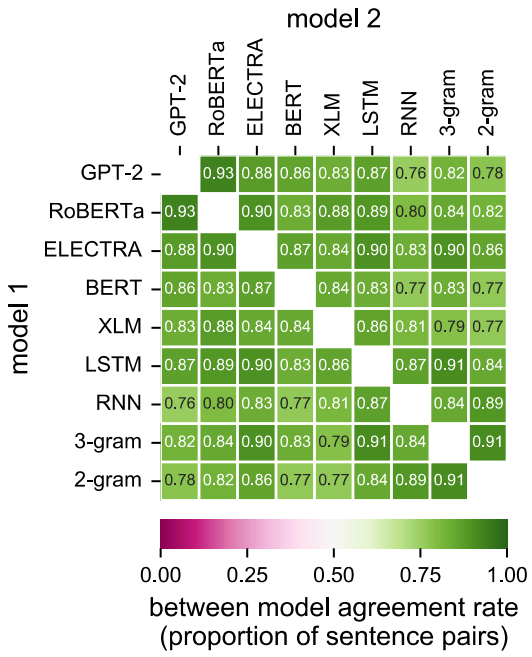

Figure S3: Between-model agreement rate on the probability ranking of the 90 randomly sampled and paired natural sentence pairs evaluated in the experiment. Each cell represents the proportion of sentence pairs for which two models make congruent probability ranking (i.e., both models assign a higher probability to sentence 1, or both models assign a higher probability to sentence 2).

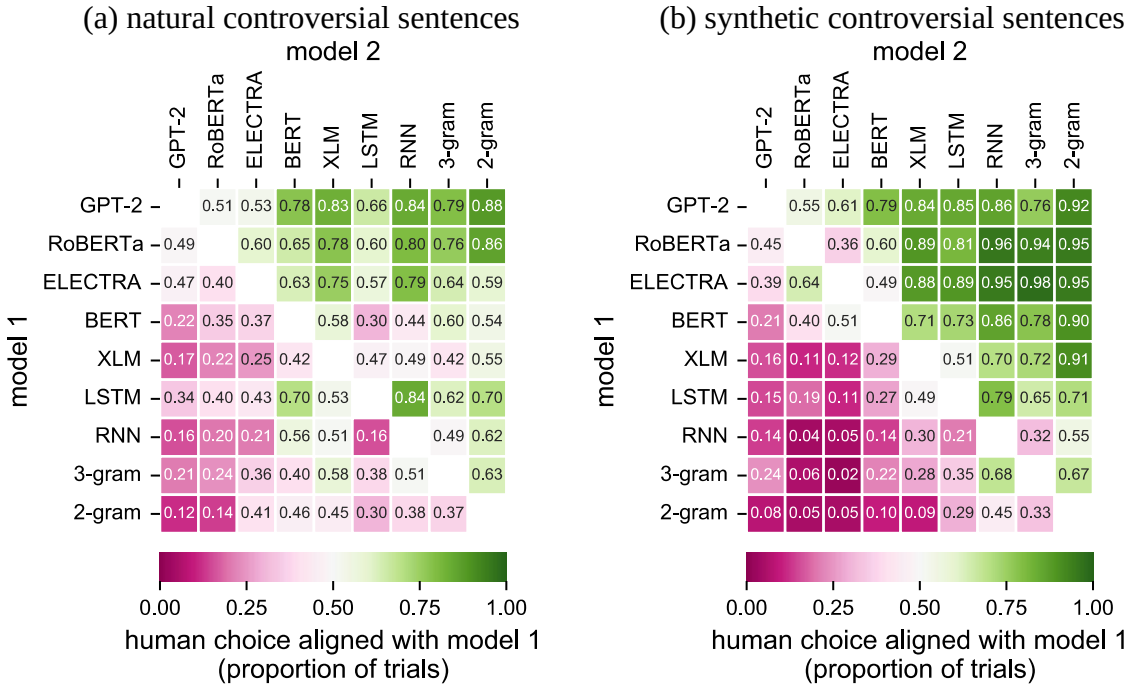

Figure S4: Pairwise model comparison of model-human consistency. For each pair of models (represented as one cell in the matrices above), the only trials considered were those in which the stimuli were either selected (a) or synthesized (b) to contrast the predictions of the two models.

For these trials, the two models always made controversial predictions (i.e., one sentence is preferred by the first model and the other sentence is preferred by the second model). The matrices above depict the proportion of trials in which the binarized human judgments aligned with the row model ("model 1"). For example, GPT-2 (top-row) was always more aligned (green hues) with the human choices than its rival models. In contrast, 2-gram (bottom-row) was always less aligned (purple hues) with the human choices than its rival models.

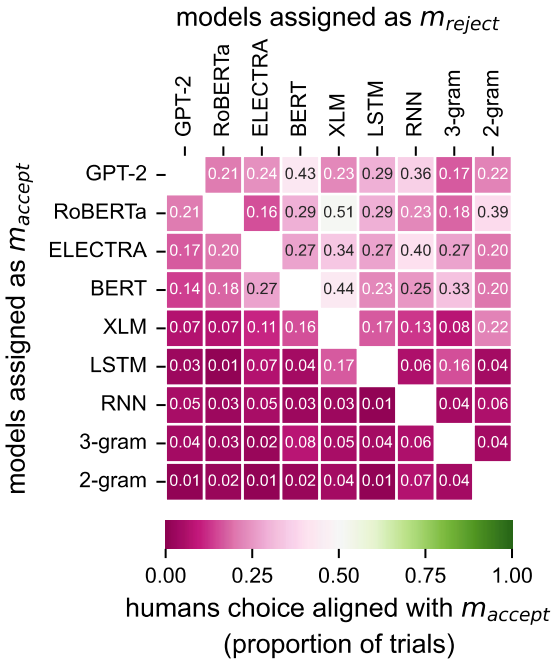

Figure S5: Pairwise model analysis of human response for natural vs. synthetic sentence pairs. In each optimization condition, a synthetic sentence *s* was formed by modifying a natural sentence *n* so the synthetic sentence would be "rejected" by one model (m_{reject}, columns), minimizing $p(s \mid m_{reject})$, and would be "accepted" by another model (m_{accept}, rows), satisfying the constraint $p(s \mid m_{accept}) \geq p(n \mid m_{accept})$. Each cell above summarizes model-human agreement in trials resulting from one such optimization condition. The color of each cell denotes the proportion of trials in which humans judged a synthetic sentence to be more likely than its natural counterpart and hence aligned with m_{accept}. For example, the top-right cell depicts human judgments for sentence pairs formed to minimize the probability assigned to the synthetic sentence by the simple 2-gram model while ensuring that GPT-2 would judge the synthetic sentence to be at least as likely as the natural sentence; humans favored the synthetic sentence in only 22 out the 100 sentence pairs in this condition.

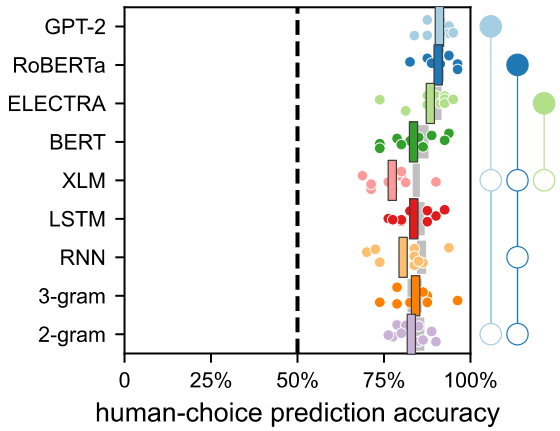

Figure S6: Model prediction accuracy for pairs of natural and synthetic sentences, evaluating each model across all of the sentence pairs in which it was targeted to rate the synthetic sentence to be less probable than the natural sentence. The data binning applied here is complementary to the one used in Fig. 1.3b, where each model was evaluated across all of the sentence pairs in which it was targeted to rate the synthetic sentence to be *at least as probable* as the natural sentence. Unlike Fig. 1.3b, where all of the models performed poorly, here no models were found to be significantly below the lower bound on the noise ceiling; typically, when a sentence was optimized to decrease its probability under any model (despite the sentence probability not decreasing under a second model), humans agreed that the sentence became less probable.

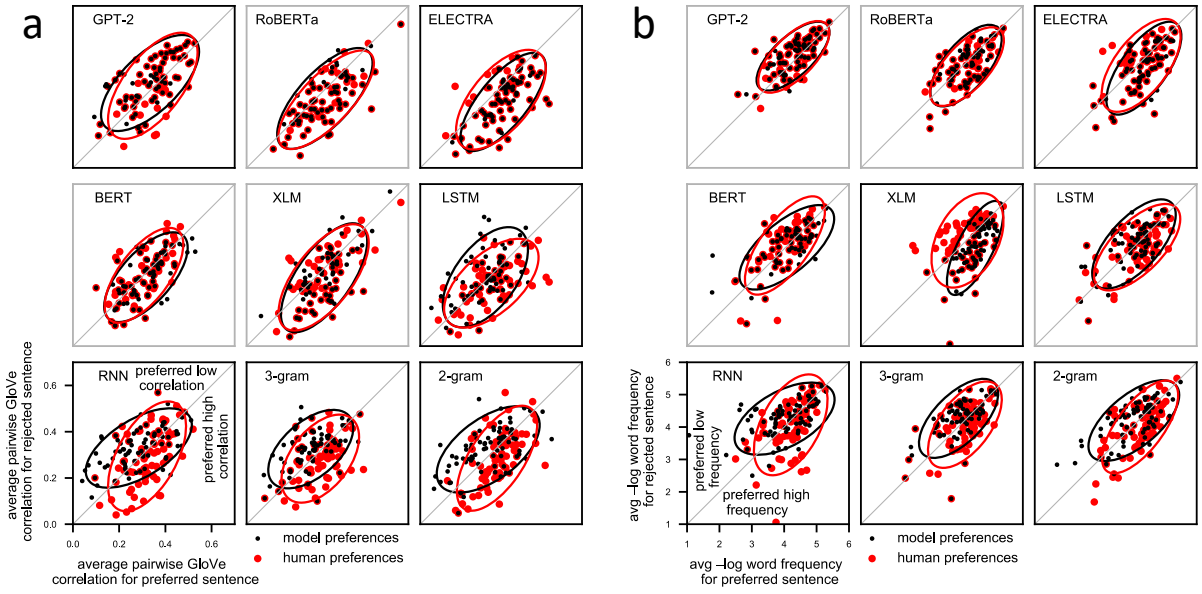

Figure S7: Linguistic feature values for synthetic sentence pairs. (a) GloVe correlation values of the preferred and rejected sentence for each synthetic sentence pair. Each panel depicts preferences for both humans (red) and a specific model (black), for sentence pairs that this model was involved in synthesizing. Black sub-panel outlines indicate significant differences between the preferences of models and humans on that particular set of sentence pairs, according to a

paired sample t-test (controlling for false discovery rate across all nine models at $q < .05$**). (b)** Same as (a), but for average log-transformed word frequency.

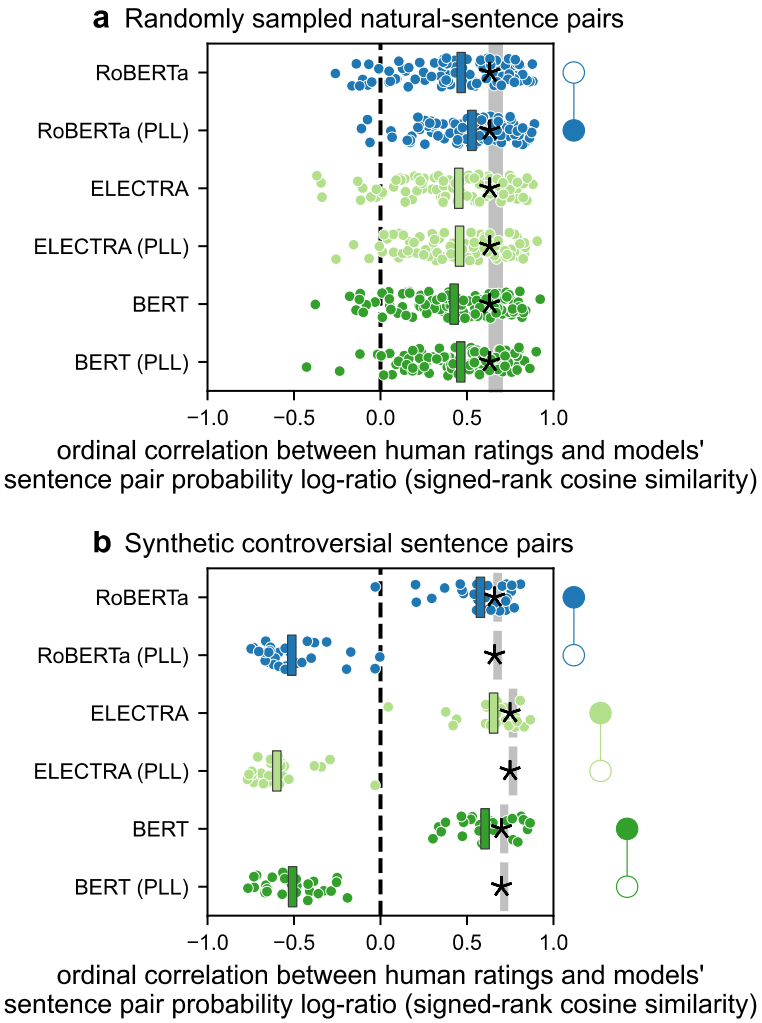

Figure S8: Human consistency of bidirectional transformers: approximate log-likelihood versus pseudo-log-likelihood (PLL). Each dot in the plots above depicts the ordinal correlation between the judgments of one participant and the predictions of one model. (a) The performance of BERT, RoBERTa, and ELECTRA in predicting the human judgments of randomly sampled natural sentence pairs in the main experiment, using two different likelihood measures: our novel approximate likelihood method (i.e., averaging multiple conditional probability chains, see Methods) and pseudo-likelihood (PLL, summating the probability of each word given all of the other words (Wang & Cho, 2019)). For each model, we statistically compared the two likelihood measures to each other and to the noise ceiling using a two-sided Wilcoxon signed-rank test across the participants. False discovery rate was controlled at $q < 0.05$ for the 9 comparisons. **When predicting human preferences of natural sentences, the pseudolog-likelihood measure is at least as accurate as our proposed approximate log-likelihood measure. (b)** Results from a follow-up experiment, in which we synthesized synthetic sentence pairs for each of the model pairs, pitting the two alternative likelihood measures against each other. Statistical testing was conducted in the same fashion as in panel a. These results indicate that for each of

the three bidirectional language models, the approximate log-likelihood measure is considerably and significantly ($q < 0.05$) more human-consistent than the pseudo-likelihood measure**. Synthetic controversial sentence pairs uncover a dramatic failure mode of the pseudo-log-likelihood measure, which remains covert when the evaluation is limited to randomly-sampled natural sentences.** See Table S1 for synthetic sentence pair examples.

sentence	pseudo-log-likelihood (PLL)	approximate log probability	# human choices
s_1: I found so many in things and called.	log$p$(s_1|BERT (PLL)) =−55.14	log$p$(s_1|BERT) =−**55.89**	**30**
s_2: Khrushchev schizophrenic so far disproportionately goldfish fished alone.	log$p$(s_2|BERT (PLL)) =−**22.84**	log$p$(s_2|BERT) =−162.31	0
s_1: Figures out if you are on the lead.	log$p$(s_1|BERT (PLL)) =−38.11	log$p$(s_1|BERT) =−**51.27**	**30**
s_2: Neighbours unsatisfactory indistinguishable misinterpreting schizophrenic on homecoming cheerleading.	log$p$(s_2|BERT (PLL)) =−**16.43**	log$p$(s_2|BERT) =−258.91	0
s_1: I just say this and not the point.	log$p$(s_1|ELECTRA (PLL)) =−34.41	log$p$(s_1|ELECTRA) =−**33.80**	**30**
s_2: Glastonbury reliably mobilize disenfranchised homosexuals underestimate unhealthy skeptics.	log$p$(s_2|ELECTRA (PLL)) =−**11.81**	log$p$(s_2|ELECTRA) =−162.62	0
s_1: And diplomacy is more people to the place.	log$p$(s_1|ELECTRA (PLL)) =−62.81	log$p$(s_1|ELECTRA) =−**47.33**	**30**
s_2: Brezhnev ingenuity disembarking Acapulco methamphetamine arthropods unaccompanied Khrushchev.	log$p$(s_2|ELECTRA (PLL)) =−**34.00**	log$p$(s_2|ELECTRA) =−230.97	0
s_1: Sometimes what looks and feels real to you.	log$p$(s_1|RoBERTa (PLL)) =−36.58	log$p$(s_1|RoBERTa) =−**51.61**	**30**
s_2: Buying something breathes or crawls aesthetically to decorate.	log$p$(s_2|RoBERTa (PLL)) =−**9.78**	log$p$(s_2|RoBERTa) =−110.27	0
s_1: In most other high priority packages were affected.	log$p$(s_1|RoBERTa (PLL)) =−71.13	log$p$(s_1|RoBERTa) =−**61.60**	**30**
s_2: Stravinsky cupboard nanny contented burglar babysitting unsupervised bathtub.	log$p$(s_2|RoBERTa (PLL)) =−**21.86**	log$p$(s_2|RoBERTa) =−164.70	0

Table S1: Examples of controversial synthetic-sentence pairs that maximally contributed to the prediction error of bidirectional transformers using pseudo-log-likelihood (PLL). For each bidirectional model, the table displays two sentence pairs on which the model failed severely when its prediction was based on pseudo-log-likelihood (PLL) estimates (Wang & Cho, 2019). In each of these sentence pairs, the PLL estimate favors sentence s_2 (higher PLL bolded), while the approximate log-likelihood estimate and most of the human subjects presented with that sentence pair preferred sentence s_1. (When more than one sentence pair induced an equal maximal error in a model, the example included in the table was chosen at random.) **Sentences with long, multi-token words (e.g., "methamphetamine") have high PLL estimates since each of their tokens is well predicted by the others tokens. And yet, the entire sentence is improbable according to human judgments and approximate log-probability estimates based on proper conditional probability chains.**

model	accepted sentence has more tokens	equal token-counts	rejected sentence has more tokens	p-value
GPT-2	24	13	3	<0.0001
RoBERTa	6	18	16	0.0656
ELECTRA	12	21	7	0.3593
BERT	4	8	28	<0.0001
XLM	2	16	22	<0.0001

Table S2: Token count control analysis. For each transformer model, we considered synthetic controversial sentence pairs where the other targeted model was also a transformer (a total of 40 sentence pairs per model). For each such pair, we evaluated the token count of the synthetic sentence to which the model assigned a higher probability ("accepted sentence") and the token count of the synthetic sentence to which the model assigned a lower probability ("rejected sentence"). For each model, this table presents the number of sentence pairs in which the accepted sentence had a higher token count, both sentences had an equal number of tokens, and the rejected sentence had a higher token count. We compared the prevalence of higher token counts in accepted and rejected sentences using a binomial test ($H0 : \pi = 0.5$) controlled for False Discovery Rate across five comparisons. GPT-2 assigned significantly more tokens to accepted sentences, whereas BERT and XLM assigned significantly more tokens to rejected sentences. RoBeRTa and ELECTRA did not show a significant difference. Note that a significant difference for a particular model does not necessarily indicate that token count biases the model's sentence probability estimates: The difference might reflect biases of the alternative models pitted against that model. **Overall, these results indicate that while certain models' probability estimates might be biased by tokenization, lower sentence probabilities were not systematically confounded by higher token counts.**