TESTING THE LIMITS OF NATURAL LANGUAGE MODELS FOR PREDICTING HUMAN LANGUAGE JUDGMENTS

Tal Golan*
Zuckerman Mind Brain Behavior Institute
Columbia University
New York, NY, USA
tal.golan@columbia.edu

Matthew Siegelman*
Department of Psychology
Columbia University
New York, NY, USA
mes2338@columbia.edu

Nikolaus Kriegeskorte
Zuckerman Mind Brain Behavior Institute,
Departments of Psychology,
Neuroscience, and Electrical Engineering
Columbia University
New York, NY, USA
n.kriegeskorte@columbia.edu

Christopher Baldassano
Department of Psychology
Columbia University
New York, NY, USA
c.baldassano@columbia.edu

ABSTRACT

Neural network language models can serve as computational hypotheses about how humans process language. We compared the model-human consistency of diverse language models using a novel experimental approach: *controversial sentence pairs*. For each controversial sentence pair, two language models disagree about which sentence is more likely to occur in natural text. Considering nine language models (including n-gram, recurrent neural networks, and transformer models), we created hundreds of such controversial sentence pairs by either selecting sentences from a corpus or synthetically optimizing sentence pairs to be highly controversial. Human subjects then provided judgments indicating for each pair which of the two sentences is more likely. Controversial sentence pairs proved highly effective at revealing model failures and identifying models that aligned most closely with human judgments. The most human-consistent model tested was GPT-2, although experiments also revealed significant shortcomings of its alignment with human perception.

Introduction

Natural language processing (NLP) models have advanced remarkably in recent years. Modeling approaches now range from simple 2-gram and 3-gram models, to recurrent neural networks [Rumelhart et al., 1986, Hochreiter and Schmidhuber, 1997], and on to transformer neural network models [Devlin et al., 2019, Liu et al., 2019, Conneau and Lample, 2019, Clark et al., 2020, Radford et al., 2019]. Within each of the neural network model classes, there exist dozens of variants that differ in their architectures, training objectives, and optimization procedures.

Cognitive scientists and NLP researchers would like to understand the ways in which these models are consistent with how humans comprehend language and in what ways models and humans diverge. A prominent approach for evaluating NLP models is handcrafted standardized benchmarks such as those in the General Language Understanding Evaluation (GLUE) [Wang et al., 2019b]. However, handcrafted benchmarks have three inherent limitations.

First, many models perform well on benchmarks such as GLUE, saturating the benchmark and hence rendering it difficult to distinguish models in terms of their alignment with humans' language understanding. Second, handcrafted tasks often evaluate language understanding from a normative perspective (i.e., requiring the model to conform to the judgments of linguists). When evaluating NLP algorithms as cognitive models, by contrast, we may be interested in identifying models that best capture language as understood by non-expert human speakers. Indeed, an estimate of the performance of human non-experts [Nangia and Bowman, 2019] ranks 20th on the GLUE leader board at the time of writing. For the successor benchmark, SuperGLUE [Wang et al., 2019a], there are already five models that score higher than human non-experts. The third, and perhaps the most concerning, disadvantage of handcrafted benchmarks is that

*The first two authors contributed equally to this work.

they can uncover only failure modes hypothesized by the designers of the benchmark. A computational model that accurately predicts human behavior in predefined benchmark tasks could still fail to match human judgments across a larger space of possible language inputs.

To avoid these three disadvantages, we propose to complement linguistics-driven handcrafted 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. We can find such sentences in large corpora of natural language. We can also synthesize novel test sentences that reveal how different models generalize beyond their training distribution.

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]. Instead, a binary forced-choice behavioral task presenting the participants with a choice between two sentences in each trial minimizes the role of between-trial context effects by setting an explicit local context within each trial.

Our experiments demonstrate that (1) it is possible to procedurally generate controversial sentence pairs for all common classes of NLP 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; (3) all current NLP model classes incorrectly assign high probability to some non-natural sentences: One can modify a natural sentence such that its probability according to a given model does not decrease but the sentence becomes markedly less probable according to human judgments. 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.

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). See the Methods section for details on the models, including their architectural properties and training tasks.

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 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. 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 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

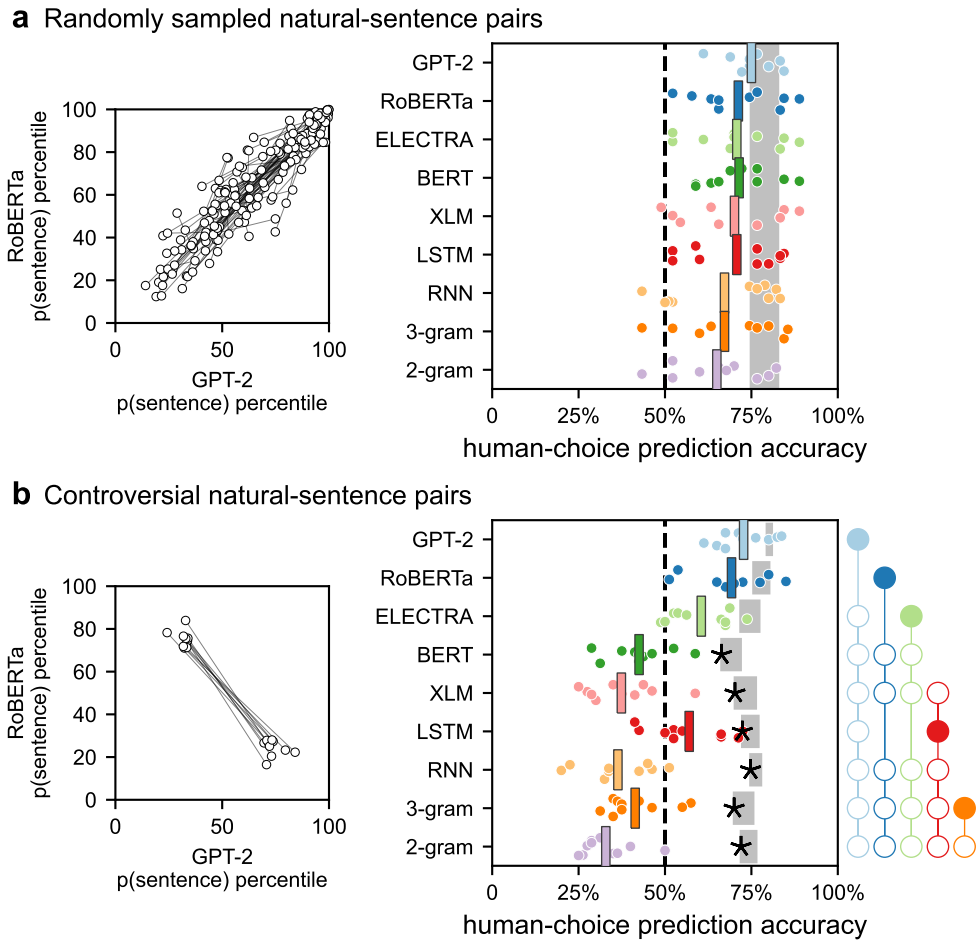


Figure 1: **Model comparison using natural sentences. (a)** (Left) Percentile-transformed sentence probabilities for GPT-2 and RoBERTa (defined relative to all sentences used in the experiment) for randomly-sampled pairs of natural sentences. Each pair of connected dots depicts one sentence pair. The 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–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 (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.

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.

Controversial natural-sentence pairs enable efficient model comparison

Next, we selected a set of controversial natural-sentence pairs by means of discrete optimization. The procedure maximized the controversiality of the resulting sentence pairs between the candidate models. For each model, we evaluated model-inferred sentence probabilities across a large corpus of 8-word Reddit sentences (Methods) and rank-transformed the probabilities separately for each model. We then selected without replacement 10 natural controversial sentence pairs for each of the 36 model pairs, yielding a total of 360 controversial sentence pairs. The optimization (Eq. 2, Methods) had two goals: (1) minimizing the rank-transformed probability of the first sentence according to the first model, subject to the constraint that this sentence must have above median probability according to the second model, and (2) minimizing the rank-transformed probability of the second sentence according to the second model, subject to the constraint that this sentence must have above median probability according to the first model. Because each sentence was to be used only once, these goals interact across the 360 sentence pairs, so the optimization procedure was implemented as a single integer programming problem for the entire experiment. By construction, this procedure yielded sentence pairs in which one sentence had a high probability rank according to one model and a low probability rank according to the other model, and the vice versa for the other sentence (for examples, see Table 1).

Each participant was presented with 36 controversial natural-sentence pairs, one per model-pair (a total of 10 controversial natural-sentence pairs were evaluated for each model pair across the 100 participants). We evaluated the models' accuracy in predicting the human choices for these sentence pairs, considering for each model only the sentence pairs in which it was one of the two targeted models (Fig. 1b). Pairwise comparisons of model prediction accuracy indicated many significant differences in terms of model-human alignment (Wilcoxon signed-rank test, controlling the false discovery rate at $q < .05$; Fig. 1b, right-hand panel). GPT-2 and RoBERTa showed the best human consistency and 2-gram the worst. A similar pattern of pairwise dominance was obtained when each comparison used only the controversial natural-sentence pairs targeting the model pair in question (Fig. S4a).

We statistically compared each model's prediction accuracy to the lower bound on the noise ceiling (the accuracy obtained by predicting each participant's responses from the other participants' responses). All models except GPT-2, RoBERTa, and ELECTRA (i.e., BERT, XLM, LSTM, RNN, 3-gram, and 2-gram) performed significantly below the noise ceiling (Wilcoxon signed-rank test, controlling the false discovery rate at $q < .05$). This result indicates

sentence	log probability (model 1)	log probability (model 2)	# human choices
n_1: Rust is generally caused by salt and sand. n_2: Where is Vernon Roche when you need him.	$\log p(n_1$|GPT-2$) = -50.72$ $\log p(n_2$|GPT-2$) = \mathbf{-32.26}$	$\log p(n_1$|ELECTRA$) = \mathbf{-38.54}$ $\log p(n_2$|ELECTRA$) = -58.26$	**10** 0
n_1: Excellent draw and an overall great smoking experience. n_2: I should be higher and tied to inflation.	$\log p(n_1$|RoBERTa$) = -67.78$ $\log p(n_2$|RoBERTa$) = \mathbf{-54.61}$	$\log p(n_1$|GPT-2$) = \mathbf{-36.76}$ $\log p(n_2$|GPT-2$) = -50.31$	**10** 0
n_1: You may try and ask on their forum. n_2: I love how they look like octopus tentacles.	$\log p(n_1$|ELECTRA$) = -51.44$ $\log p(n_2$|ELECTRA$) = \mathbf{-35.51}$	$\log p(n_1$|LSTM$) = \mathbf{-44.24}$ $\log p(n_2$|LSTM$) = -66.66$	**10** 0
n_1: Grow up and quit whining about minor inconveniences. n_2: The extra a is the correct Sanskrit pronunciation.	$\log p(n_1$|BERT$) = -82.74$ $\log p(n_2$|BERT$) = \mathbf{-51.06}$	$\log p(n_1$|GPT-2$) = \mathbf{-35.66}$ $\log p(n_2$|GPT-2$) = -51.10$	**10** 0
n_1: I like my password manager for this reason. n_2: Kind of like clan of the cave bear.	$\log p(n_1$|XLM$) = -68.93$ $\log p(n_2$|XLM$) = \mathbf{-44.24}$	$\log p(n_1$|RoBERTa$) = \mathbf{-49.61}$ $\log p(n_2$|RoBERTa$) = -67.00$	**10** 0
n_1: We have raised a Generation of Computer geeks. n_2: I mean when the refs are being sketchy.	$\log p(n_1$|LSTM$) = -66.41$ $\log p(n_2$|LSTM$) = \mathbf{-42.04}$	$\log p(n_1$|ELECTRA$) = \mathbf{-36.57}$ $\log p(n_2$|ELECTRA$) = -52.28$	**10** 0
n_1: This is getting ridiculous and ruining the hobby. n_2: I think the boys and invincible are better.	$\log p(n_1$|RNN$) = -100.65$ $\log p(n_2$|RNN$) = \mathbf{-45.16}$	$\log p(n_1$|LSTM$) = \mathbf{-43.50}$ $\log p(n_2$|LSTM$) = -59.00$	**10** 0
n_1: Then attach them with the supplied wood screws. n_2: Sounds like you were used both a dog.	$\log p(n_1$|3-gram$) = -119.09$ $\log p(n_2$|3-gram$) = \mathbf{-92.07}$	$\log p(n_1$|GPT-2$) = \mathbf{-34.84}$ $\log p(n_2$|GPT-2$) = -52.84$	**10** 0
n_1: Cream cheese with ham and onions on crackers. n_2: I may have to parallel process that drinking.	$\log p(n_1$|2-gram$) = -131.99$ $\log p(n_2$|2-gram$) = \mathbf{-109.46}$	$\log p(n_1$|RoBERTa$) = \mathbf{-54.62}$ $\log p(n_2$|RoBERTa$) = -70.69$	**10** 0

Table 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.)

a misalignment between these models' predictions and human judgments which was only revealed when using controversial sentence pairs.

Synthesizing controversial sentence pairs

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 sentence to drive different models to make opposing predictions, forming *synthetic* controversial sentences that may lay outside any natural language corpora.

We developed a procedure for synthesizing controversial sentence pairs in which naturally occurring sentences serve as initializations for synthetic sentences as well as reference points that guide sentence synthesis (Figure 2). We begin by sampling a naturally occurring sentence. We then iteratively replace words in the sentence with words from a predefined vocabulary, aiming to minimize the sentence probability assigned to the synthetic sentence by one language model, subject to the constraint that the synthetic sentence remains at least as likely as the natural sentence according to an alternative model (see Methods, "Generating Synthetic Controversial Sentence-Pairs" for further details). Conceptually, this approach resembles Maximum Differentiation (MAD) competition Wang and Simoncelli [2008], introduced to compare models of image quality assessment.

Synthetic-sentence pairs enable even greater disentanglement of model predictions

For each model pair, 10 controversial synthetic-sentence pairs were evaluated by the human subjects. 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. 3a, right-hand panel). 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% 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 (Wilcoxon signed-rank test, controlling the false discovery rate for all 36 model pairs at $q < .05$). All of the models

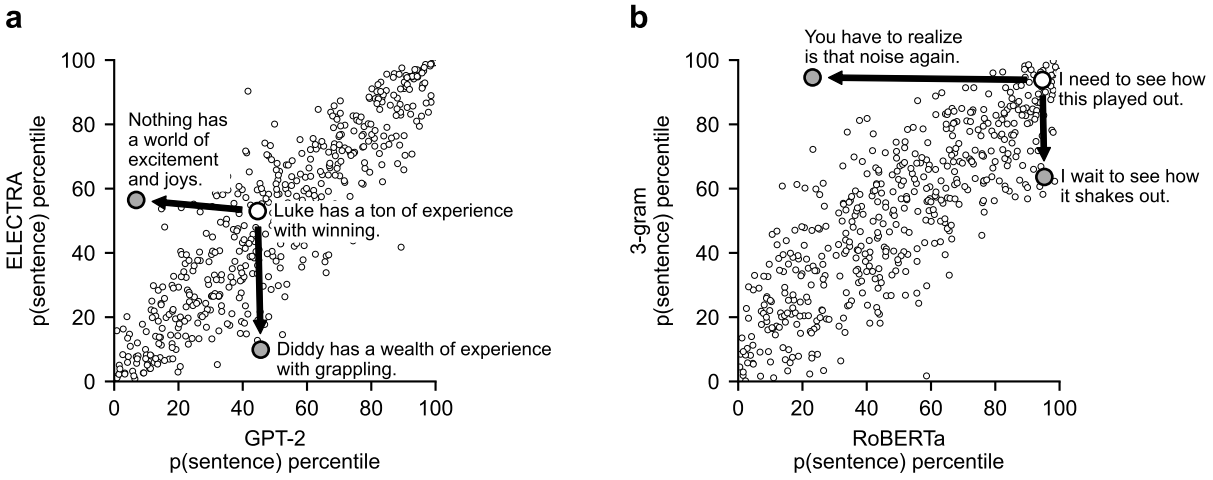


Figure 2: **Synthesizing controversial sentence pairs.** The small open dots denote 500 randomly sampled natural sentences. The big open dot denotes the natural sentence used for initializing the controversial sentence optimization, and the closed dots are the resulting synthetic sentences. **(a)** In 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".

except for GPT-2 were found to be significantly below the lower bound on the noise ceiling (Wilcoxon signed-rank test, controlling the false discovery rate for all nine models at $q < .05$). Evaluating each model pair only on the controversial synthetic-sentence pairs specifically targeting that particular model pair yielded a similar model ranking (Fig. S4b).

Examples of controversial synthetic-sentence pairs that maximally contributed to the models' prediction error appear in Table 2. Qualitatively inspecting these sentence pairs suggests that for each language model, the optimization unveiled nonsensical or even nongrammatical sentences that were strongly favored by the model over more natural alternatives.

sentence	log probability (model 1)	log probability (model 2)	# human choices
s_1: You can reach his stories on an instant. s_2: Anybody can behead a rattles an an antelope.	$\log p(s_1\|\text{GPT-2}) = -64.92$ $\log p(s_2\|\text{GPT-2}) = \mathbf{-40.45}$	$\log p(s_1\|\text{RoBERTa}) = \mathbf{-59.98}$ $\log p(s_2\|\text{RoBERTa}) = -90.87$	**10** 0
s_1: However they will still compare you to others. s_2: Why people who only give themselves to others.	$\log p(s_1\|\text{RoBERTa}) = -53.40$ $\log p(s_2\|\text{RoBERTa}) = \mathbf{-48.66}$	$\log p(s_1\|\text{GPT-2}) = \mathbf{-31.59}$ $\log p(s_2\|\text{GPT-2}) = -47.13$	**10** 0
s_1: He healed faster than any professional sports player. s_2: One gets less than a single soccer team.	$\log p(s_1\|\text{ELECTRA}) = -48.77$ $\log p(s_2\|\text{ELECTRA}) = \mathbf{-38.25}$	$\log p(s_1\|\text{BERT}) = \mathbf{-50.21}$ $\log p(s_2\|\text{BERT}) = -59.09$	**10** 0
s_1: That is the narrative we have been sold. s_2: This is the week you have been dying.	$\log p(s_1\|\text{BERT}) = -56.14$ $\log p(s_2\|\text{BERT}) = \mathbf{-50.66}$	$\log p(s_1\|\text{GPT-2}) = \mathbf{-26.31}$ $\log p(s_2\|\text{GPT-2}) = -39.50$	**10** 0
s_1: The resilience is made stronger by early adversity. s_2: Every thing is made alive by infinite Ness.	$\log p(s_1\|\text{XLM}) = -62.95$ $\log p(s_2\|\text{XLM}) = \mathbf{-42.95}$	$\log p(s_1\|\text{RoBERTa}) = \mathbf{-54.34}$ $\log p(s_2\|\text{RoBERTa}) = -75.72$	**10** 0
s_1: President Trump threatens to storm the White House. s_2: West Surrey refused to form the White House.	$\log p(s_1\|\text{LSTM}) = -58.78$ $\log p(s_2\|\text{LSTM}) = \mathbf{-40.35}$	$\log p(s_1\|\text{RoBERTa}) = \mathbf{-41.67}$ $\log p(s_2\|\text{RoBERTa}) = -67.32$	**10** 0
s_1: Las beans taste best with a mustard sauce. s_2: Roughly lanes being alive in a statement ratings.	$\log p(s_1\|\text{RNN}) = -131.62$ $\log p(s_2\|\text{RNN}) = \mathbf{-49.31}$	$\log p(s_1\|\text{RoBERTa}) = \mathbf{-60.58}$ $\log p(s_2\|\text{RoBERTa}) = -99.90$	**10** 0
s_1: You are constantly seeing people play the multi. s_2: This will probably the happiest contradicts the hypocrite.	$\log p(s_1\|\text{3-gram}) = -107.16$ $\log p(s_2\|\text{3-gram}) = \mathbf{-91.59}$	$\log p(s_1\|\text{ELECTRA}) = \mathbf{-44.79}$ $\log p(s_2\|\text{ELECTRA}) = -75.83$	**10** 0
s_1: A buyer can own a genuine product also. s_2: One versed in circumference of highschool I rambled.	$\log p(s_1\|\text{2-gram}) = -127.35$ $\log p(s_2\|\text{2-gram}) = \mathbf{-113.73}$	$\log p(s_1\|\text{ELECTRA}) = \mathbf{-40.21}$ $\log p(s_2\|\text{ELECTRA}) = -92.61$	**10** 0

Table 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 s_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 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.)

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. Twenty pairs of sentences for each model pair were presented to the participants, 10 for which the natural sentence was shown with one of the derived synthetic sentences and 10 for which the natural sentence was shown with the other derived synthetic sentence. For each of these trials, there is (by construction) a model that rates the synthetic sentence to be at least as probable as the 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. That would bring the noise ceiling and the model prediction accuracy to a similar level–50% or more. In reality, the sample of human participants we tested showed a systematic preference for the natural sentences over their synthetic counterparts (Fig. 3b, right-hand panel), even when the synthetic sentences were formed such that the stronger models (i.e., GPT-2, RoBERTA, or ELECTRA) would favor them over the natural sentences. See Table 3 for examples.

This result demonstrates an abundance of imperfections in the distribution of language as learned by even the stronger models. By forming synthetic sentences that are at least as probable as reference naturally-occurring sentences but are improbable according to an alternative model, we find stark inconsistencies between language models and humans. Evaluating natural sentence preference separately for each model-pairing (Fig. S5) demonstrates that the alternative model targeted to "reject" the synthetic sentence (i.e., rate it to be less probable than the natural sentence) does not have to be particularly strong. For example, when GPT-2 is targeted to "accept" the synthetic sentence (i.e., assign it with at least as high probability as the natural sentence) and the 2-gram or another weak model is targeted to reject the synthetic sentence, the resulting synthetic sentences are still viewed as less probable than the natural sentences (top row of Fig. S5).

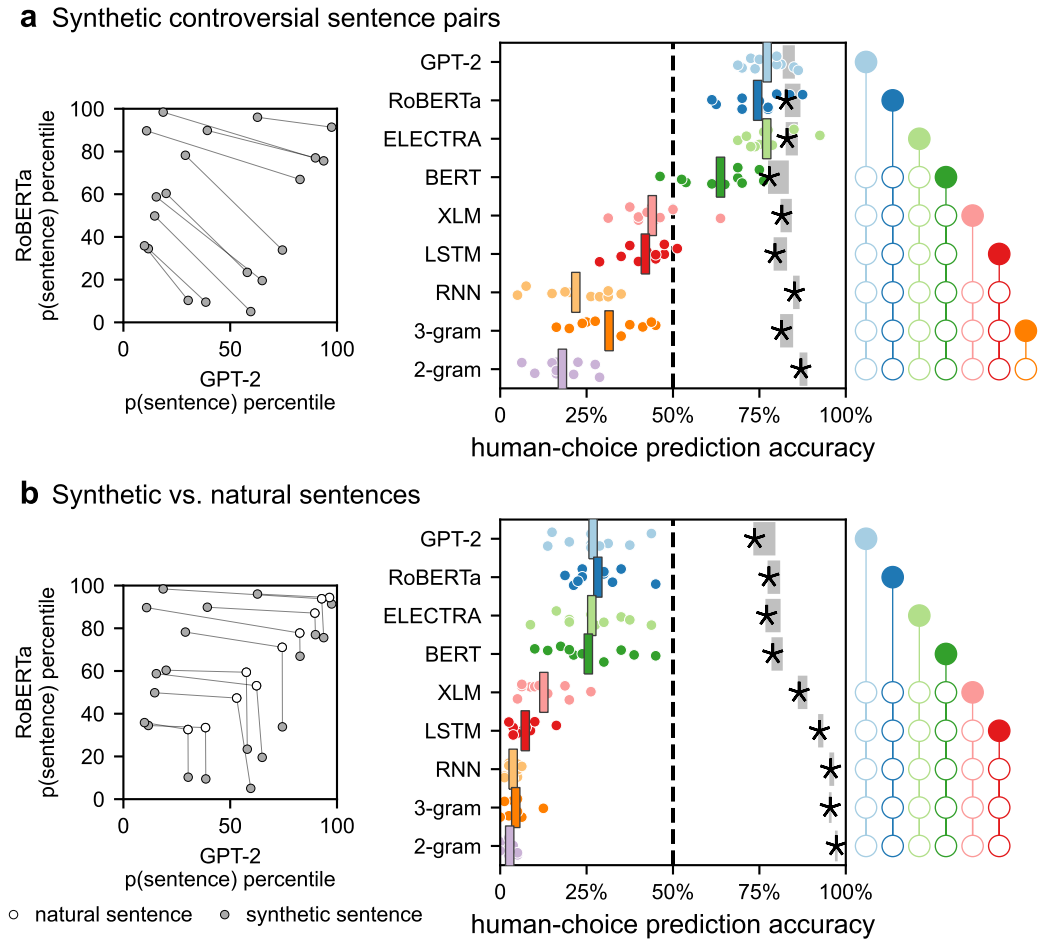


Figure 3: **Model comparison using synthetic sentences. (a)** (Left) Percentile-transformed sentence probabilities for GPT-2 and RoBERTa for controversial synthetic-sentence pairs. Each pair of connected dots depict one sentence pair. (Right) Model prediction accuracy, measured as the proportion of trials 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 (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–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 Figure 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. s: We either wish for it or ourselves do.	$\log p(n\|\text{GPT-2}) = -36.46$ $\log p(s\|\text{GPT-2}) = \mathbf{-36.15}$	$\log p(n\|\text{2-gram}) = \mathbf{-106.95}$ $\log p(s\|\text{2-gram}) = -122.28$	**10** 0
n: This is why I will never understand boys. s: This is why I will never kiss boys.	$\log p(n\|\text{RoBERTa}) = -46.88$ $\log p(s\|\text{RoBERTa}) = \mathbf{-46.75}$	$\log p(n\|\text{2-gram}) = \mathbf{-103.11}$ $\log p(s\|\text{2-gram}) = -107.91$	**10** 0
n: One of the ones I did required it. s: Many of the years I did done so.	$\log p(n\|\text{ELECTRA}) = -35.97$ $\log p(s\|\text{ELECTRA}) = \mathbf{-35.77}$	$\log p(n\|\text{LSTM}) = \mathbf{-40.89}$ $\log p(s\|\text{LSTM}) = -46.25$	**10** 0
n: There were no guns in the Bronze Age. s: There is rich finds from the Bronze Age.	$\log p(n\|\text{BERT}) = -48.48$ $\log p(s\|\text{BERT}) = \mathbf{-48.46}$	$\log p(n\|\text{ELECTRA}) = \mathbf{-30.40}$ $\log p(s\|\text{ELECTRA}) = -44.34$	**10** 0
n: You did a great job on cleaning them. s: She did a great job at do me.	$\log p(n\|\text{XLM}) = -40.38$ $\log p(s\|\text{XLM}) = \mathbf{-39.89}$	$\log p(n\|\text{RNN}) = \mathbf{-43.47}$ $\log p(s\|\text{RNN}) = -61.03$	**10** 0
n: This logic has always seemed flawed to me. s: His cell has always seemed instinctively to me.	$\log p(n\|\text{LSTM}) = -39.77$ $\log p(s\|\text{LSTM}) = \mathbf{-38.89}$	$\log p(n\|\text{RNN}) = \mathbf{-45.92}$ $\log p(s\|\text{RNN}) = -62.81$	**10** 0
s: Stand near the cafe and sip your coffee. n: Sit at the front and break your neck.	$\log p(s\|\text{RNN}) = -65.55$ $\log p(n\|\text{RNN}) = \mathbf{-44.18}$	$\log p(s\|\text{ELECTRA}) = \mathbf{-34.46}$ $\log p(n\|\text{ELECTRA}) = -34.65$	**10** 0
n: Most of my jobs have been like this. s: One of my boyfriend have been like this.	$\log p(n\|\text{3-gram}) = -80.72$ $\log p(s\|\text{3-gram}) = \mathbf{-80.63}$	$\log p(n\|\text{LSTM}) = \mathbf{-35.07}$ $\log p(s\|\text{LSTM}) = -41.44$	**10** 0
n: They even mentioned that I offer white flowers. s: But even fancied that would logically contradictory philosophies.	$\log p(n\|\text{2-gram}) = -113.38$ $\log p(s\|\text{2-gram}) = \mathbf{-113.24}$	$\log p(n\|\text{BERT}) = \mathbf{-62.81}$ $\log p(s\|\text{BERT}) = -117.98$	**10** 0

Table 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 (Figure 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 ($q < .05$ for all of the comparisons marked in Fig. 4). All of the models (including GPT-2) were found to be significantly less accurate than the lower bound on the noise ceiling (Wilcoxon signed-rank test, controlling the false discovery rate for nine comparisons at $q < .05$).

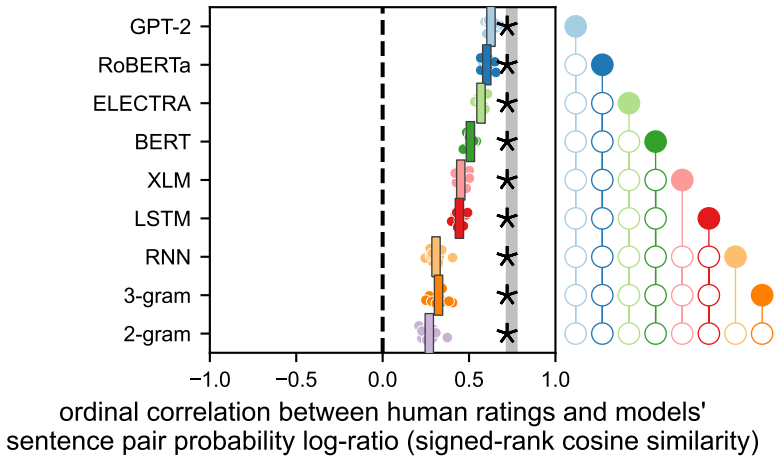


Figure 4: **Ordinal correlation of the models' sentence probability log-ratios and human Likert ratings.** For each sentence pair, model prediction was quantified by $\log \frac{p(s^1|m)}{p(s^2|m)}$. This log-ratio was correlated with the Likert ratings of each particular participant, using signed-rank 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 Figure 1's caption for details on the visualization convention.

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 semantic coherence of each sentence (measured by the average pairwise correlation between semantic GloVe vector representations [Pennington et al., 2014] of all eight words) and the average log-transformed word frequency for each sentence (extracted from the publicly available subtlex database [Van Heuven et al., 2014]). 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 low-coherence sentences, while ELECTRA was significantly more likely to prefer high-coherence sentences (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.

Discussion

In this study, we introduced a model-driven experimental approach to comparing candidate computational models of human language processing. Specifically, we probed the models' ability 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. This approach allowed us to statistically distinguish between language models in terms of their human consistency and test how well the models' predictions generalize beyond the training distribution.

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 word prediction task) were the most predictive of human judgments on controversial natural-sentence pairs (Fig. 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. 3a); and (3) GPT-2 was the most human-consistent model when considering the entire behavioral dataset we collected (Fig. 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. 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 [e.g., Wehbe et al., 2014, Toneva and 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 and King, 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], as well as with evidence from model-based prediction of human reading speed [Wilcox et al., 2021, Merkx and Frank, 2021] and N400 amplitude [Merkx and 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. GPT-2 performed the best overall, consistent with this model's advantage in predicting ECoG and fMRI responses to natural language [Schrimpf et al., 2021; see also Goldstein et al., 2022, Caucheteux and King, 2022].

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. 3b) demonstrates that GPT-2 does not fully emulate the computations employed in human processing of even short sentences. This outcome is 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 mapped GPT-2 hidden-layer activations to brain data by means of regularized linear regression (i.e., a linearized encoding analysis was employed). This learned projection into the lower-dimensional space of neural responses identifies a subspace within GPT-2's language representation that is well-aligned with brain responses, but does not necessarily imply that GPT-2's overall sentence probabilities will be 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 testing model-brain correspondence with controversial synthetic sentences designed to elicit distinct representations in different language models will reveal considerable discrepancies between GPT-2 (as well as the other off-the-shelf language models) and human neural representations. In addition to our approach for optimizing controversiality across sentences, complementary methods could be used for generating individual sentences as stimuli for functional MRI experiments, for example with controversial transition probabilities between consecutive words [Rakocevic, 2021].

Testing with controversial sentences can be seen as a generalization test of language models: having been trained only on typical sentences, to what extent can they judge the probability of atypical sentences? For humans, comprehending strange-sounding language with atypical constructions or unapparent meaning is not a cognitive challenge, as pragmatic judgments can be made about a speaker's intentions from environmental and linguistic context [Goodman and Frank, 2016]. While the most advanced language models also utilize linguistic context to evaluate individual sentences, perturbations in typical syntactic or semantic constructions appear to have a much more dramatic effect on their outputs. The fact that even the best model we considered, GPT-2, failed this generalization test can be interpreted in two ways. First, it might be that the next-word prediction task is indeed sufficient for achieving a fully human aligned model but GPT-2 still falls short in terms its architecture, learning rules, or training data. If this is the case, stronger next-word prediction models (i.e., GPT-3 and its future predecessors) might close the gap between models and humans. Alternatively, it might be that 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 NLP 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 NLP models has motivated proposed generalizations of adversarial methods to textual inputs [for review, see 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 NLP models have instead relied on heuristic constraints aiming to limit the adversarial text modifications from altering the meaning of the text as understood by human speakers. Some of these adversarial methods introduce only low-level replacements that are unlikely to affect semantics, for example, flipping a character [Liang et al., 2018, Ebrahimi et al., 2018]. Others seek to introduce higher-level but semantics-preserving modifications, such as changing number or gender [Abdou et al., 2020] or replacing words with their synonyms [Alzantot et al., 2018, Ribeiro et al., 2018, Ren et al., 2019]. However, since these heuristics rely on rough approximations of human language processing, none of them is guaranteed to result in genuine adversarial examples (i.e., inputs that induce an incorrect response in the model, as judged against human evaluation), many of these methods indeed introduce modifications that fail to preserve semantics [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 NLP models.

By contrast, testing NLP 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 it differs from red-teaming in two respects: (1) Here, the roles of the models are symmetric (there is no "attacking" and "attacked" models) and (2) Importantly, the controversiality objective requires no filtering of the generated examples for their adversarial status.

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.

First, not all available models were tested in our experiment, for a number of reasons. Primarily, the space of language models is constantly expanding as small improvements are made to existing models to yield slightly higher performance accuracy on some key task. Our model selection reflects a selection of popular models available when the time the stimulus set was created. Secondly, at the time of conducting the study, some models of interest [e.g. GPT-3, Brown et al., 2020] were not available for full inspection and analysis but instead were offered only through a limited API. And last, we were limited by the number of models we could include in this study, given that the number of trials required to test each pair of models would increase multiplicatively with each new model, and we aimed to include multiple models from each of the three classes. Future work can introduce (potentially adaptive) controversial sentence optimization procedures that consider large sets of candidate models, allowing for greater model coverage than our simpler pairwise approach.

A more substantial limitation of the current study is that we measured only the sentence preferences of human participants, and so cannot directly assess whether these judgments were driven by semantic, grammatical, or other linguistic information. Similarly, we did not optimize sentences to be controversial with respect to specific linguistic features, as in handcrafted datasets like BLiMP [Warstadt et al., 2020] with sentence pairs designed to probe particular aspects of linguistic acceptability. Our post-hoc analysis of low-level linguistic features suggested that some models tend to differ from humans in the influence of low-level linguistic criteria on the evaluation of sentence probabilities, but it is not clear whether this contrast represents deep inductive biases resulting from differences in model architecture and training. In future work these kinds of features could be explicitly controlled as part of the optimization process. For example, sentiment classifiers could be trained on BERT and GPT-2 representations, and a sentence could be optimized to have a strong positive sentiment according to BERT and a strong negative sentiment according to GPT-2 (while still having a typical sentence probability according to both BERT and GPT-2).

Further extensions of our work could explore alternative methods for optimizing controversial sentences. Our current optimization procedure could be made more effective by replacing the simple hill-climbing search method we used with less local search procedures, such as beam search or evolutionary algorithms. A more substantive change would be to reformulate the optimization objective based on an Information-Theoretic Optimal Experimental Design approach [Lindley, 1956]. Specifically, if language model probabilities can be mapped to calibrated human judgment probabilities, sentence pairs can be synthesized to maximize the expected information gain with respect to the model comparison task. However, our results suggest that at least for the current language models, this full Information-Theoretic formulation may not be necessary; our current approach (focusing on sentence rank-ordering rather than calibrated probabilities) delineates a clear hierarchy of models based on empirical model-human alignment and uncovers stark discrepancies between all of the candidate language models and human judgments.

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].

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) w_1. In 3-gram models, all unique three-word phrases are counted, and upcoming word probabilities (probability of w_3 conditioned on previous words w_1 and 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: i) public comments from the social media website Reddit (`reddit.com`) acquired using the public API at `pushshift.io`, ii) articles from Wikipedia, iii) English books and poetry available for free at Project Gutenberg (`gutenberg.org`), and iv) 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 and 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 and 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 and 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: i) 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 ii) 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. 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: i) the same MLM task used in both BERT and RoBERTa, ii) a causal language modeling task where upcoming words are predicted from left to right, and iii) 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.

Evaluating sentence-level probabilities

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 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. Some attempts have been made to evaluate sentence probabilities from bidirectional transformer models (e.g., Wang and Cho [2019]), but there is no clear consensus on the appropriate method for doing so. 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 word positions as possible construction orders. 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 (analogous to the random word visitation order used to sample serial reproduction chains, Yamakoshi et al. [2022]) 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:

$$\begin{aligned}\log p(w = \text{beehive}) =& 0.5\big(\log p(w_1 = \text{bee} \mid w_2 = \text{MASK}) + \log p(w_2 = \text{hive} \mid w_1 = \text{bee})\big) \\ &+ 0.5\big(\log p(w_2 = \text{hive} \mid w_1 = \text{MASK}) + \log p(w_1 = \text{bee} \mid w_2 = \text{hive})\big)\end{aligned} \tag{1}$$

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 $27{,}079$ words included in the training corpus for the n-gram and recurrent neural network models, and to exclude those included in a predetermined list of inappropriate words and phrases.

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 $\binom{85749}{2} \approx 3.67 \times 10^9$ possible sentence pairs can be formed.

We aimed to select 360 controversial sentence pairs, devoting 10 sentence pairs to each of the 36 models pairs. First, we defined two 360-long integer vectors $\boldsymbol{m}^1$ and $\boldsymbol{m}^2$, specifying for each of the 360 yet unselected sentence pairs which model pair they contrast. We then selected 360 sentence pairs $\left(s_1^1, s_1^2\right), \left(s_2^1, s_2^2\right)..., \left(s_{360}^1, s_{360}^2\right)$ by solving the following minimization problem:

$$\{(s_j^{1^*}, s_j^{2^*}) \mid j = 1, 2, ..360\} = \underset{\boldsymbol{s^1}, \boldsymbol{s^2}}{\operatorname{argmin}} \quad \sum_j \left(r(s_j^1|m_j^1) + r(s_j^2|m_j^2)\right) \tag{2}$$

$$\text{subject to} \quad \forall_j r(s_j^1|m_j^2) \geq 0.5 \tag{2a}$$

$$\forall_j r(s_j^2|m_j^1) \geq 0.5 \tag{2b}$$

$$\text{All 720 sentences are unique.} \tag{2c}$$

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 $\times$ 360 $\times$ 2 (sentences, trials, pair members) and the fractional sentence probabilities ranks as a matrix R of dimensions 85,749 $\times$ 9 (sentences, models). This enabled us to express and solve the selection problem in Eq. 2 as a standard ILP problem:

$$\boldsymbol{S}^* = \underset{\boldsymbol{S}}{\operatorname{argmin}} \sum_{i,j} \quad \boldsymbol{S}_{i,j,1}\boldsymbol{R}_{i,m_j^1} + \boldsymbol{S}_{i,j,2}\boldsymbol{R}_{i,m_j^2} \tag{3}$$

$$\text{subject to} \quad \boldsymbol{S}_{i,j,1}\boldsymbol{R}_{i,m_j^2} \geq 0.5 \tag{3a}$$

$$\boldsymbol{S}_{i,j,2}\boldsymbol{R}_{i,m_j^1} \geq 0.5 \tag{3b}$$

$$\forall_i \sum_{j,k} \boldsymbol{S}_{i,j,k} \leq 1 \text{ (each sentence } i \text{ is used only once in the experiment)} \tag{3c}$$

$$\left.\begin{array}{l} \forall_j \sum_i \boldsymbol{S}_{i,j,1} = 1 \\ \forall_j \sum_i \boldsymbol{S}_{i,j,2} = 1 \end{array}\right\} \text{(each trial } j \text{ is allocated exactly one sentence pair)} \tag{3d}$$

$$\boldsymbol{S} \text{ is binary} \tag{3e}$$

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. Each synthetic sentence was generated as a solution for a constrained minimization problem:

$$\begin{aligned} s^* = \underset{s}{\operatorname{argmin}} & \log p(s \mid m_{reject}) \\ \text{subject to } & \log p(s \mid m_{accept}) \geq \log p(n \mid m_{accept}) \end{aligned} \tag{4}$$

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 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.

When evaluating potential replacement words, we only considered words that were sufficiently prevalent and hence expected to be present across training corpora. This vocabulary was determined by intersecting the list of words in the subtlex database [Van Heuven et al., 2014] with the corpus used to train the n-gram and recurrent neural network models, which includes about 300M words. We excluded all words occurring less than 300 times in the latter corpus (i.e., less frequent than one in a million). This resulted in a list of $25{,}258$ potential replacement words.

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 $23{,}180$ sentences resulting from each of the $23{,}180$ 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, 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:

$$\log \hat{p}(s\{i\} \leftarrow w \mid m) = \beta_1 \log p(s\{i\} = w \mid m, si \leftarrow mask) + \beta_0 \tag{5}$$

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 bi-directional) was used to predict $\log p(s \mid m_{reject})$ for each of the $23{,}180$ 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. 5) 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

$$c_{m_1,m_2}(n, s_1, s_2) = \log \frac{p(n \mid m_1)}{p(s_1 \mid m_1)} + \log \frac{p(n \mid m_2)}{p(s_2 \mid m_2)}, \tag{6}$$

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). 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.

Evaluation of model-human consistency: binarized-judgments

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. 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 by a Wilcoxon signed-rank test using these 10 independent accuracy outcomes as paired samples.

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. Models' accuracy was tested against this lower bound by a Wilcoxon signed-rank test. 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.

For each analysis, the false discovery rate across multiple comparisons was controlled by the Benjamini-Hochberg procedure [Benjamini and Hochberg, 1995].

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:

$$LR(s_i^1, s_i^2 \mid m) = \log \frac{p(s_i^2 \mid m)}{p(s_i^1 \mid m)} \tag{7}$$

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:

$$r(s_i^1, s_i^2 \mid h) = \text{sign}(y_0(s_i^1, s_i^2 \mid h)) \cdot R\big(\,\big|y_0(s_i^1, s_i^2 \mid h)\big|\,\big), \tag{8}$$

where $y_0(s_i^1, s_i^2 \mid h))$ 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:

$$r(s_i^1, s_i^2 \mid m) = \text{sign}(LR(s_i^1, s_i^2 \mid m)) \cdot R\big(\,\big|LR(s_i^1, s_i^2 \mid m)\big|\,\big), \tag{9}$$

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):

$$S_{CSR} = \frac{\sum_i r(s_i^1, s_i^2 \mid m) r(s_i^1, s_i^2 \mid h)}{\sqrt{\sum_i r(s_i^1, s_i^2 \mid m)^2}\sqrt{\sum_i r(s_i^1, s_i^2 \mid h)^2}}. \tag{10}$$

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:

$$\mathbb{E}(S_{CSR}) = \frac{\sum_i \mathbb{E}\big(r(s_i^1, s_i^2 \mid m)\big)\mathbb{E}\big(r(s_i^1, s_i^2 \mid h)\big)}{\sqrt{\sum_{k=1}^n k^2}\sqrt{\sum_{k=1}^n k^2}} = \frac{\sum_i \bar{r}(s_i^1, s_i^2 \mid m)\bar{r}(s_i^1, s_i^2 \mid h)}{\sum_{k=1}^n k^2}, \tag{11}$$

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 Schütt et al. [2021], section 5.1.2, for a modern treatment). We empirically confirmed that averaging S_{CSR} as defined in Eq. 10 across a large number of random tie-breaking draws converges to $\mathbb{E}(S_{CSR})$ 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 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 ($\bar{\mathbf{r}}/\sqrt{\sum k^2}$) 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).

Data and code availability

Sentence optimization and data analysis code, experimental stimuli, and detailed behavioral testing results are available at github.com/dpmlab/contstimlang.

Acknowledgements

This publication was made possible with the support of the Charles H. Revson Foundation to TG. The statements made and views expressed, however, are solely the responsibility of the authors.

References

Mostafa Abdou, Vinit Ravishankar, Maria Barrett, Yonatan Belinkov, Desmond Elliott, and Anders Søgaard. The sensitivity of language models and humans to Winograd schema perturbations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7590–7604, Online, July 2020. Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.679. URL `https://aclanthology.org/2020.acl-main.679`.

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural*

Language Processing, pages 2890–2896, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi:10.18653/v1/D18-1316. URL https://aclanthology.org/D18-1316.

Yoav Benjamini and Yosef Hochberg. 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, 1995. doi:https://doi.org/10.1111/j.2517-6161.1995.tb02031.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1995.tb02031.x.

Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1):134, December 2022. ISSN 2399-3642. doi:10.1038/s42003-022-03036-1. URL https://www.nature.com/articles/s42003-022-03036-1.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=r1xMH1BtvB.

Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/c04c19c2c2474dbf5f7ac4372c5b9af1-Paper.pdf.

David V. Cross. Sequential dependencies and regression in psychophysical judgments. *Perception & Psychophysics*, 14(3):547–552, October 1973. ISSN 0031-5117, 1532-5962. doi:10.3758/BF03211196. URL http://link.springer.com/10.3758/BF03211196.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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)*, pages 4171–4186. Association for Computational Linguistics, 2019. doi:10.18653/v1/n19-1423. URL https://doi.org/10.18653/v1/n19-1423.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. HotFlip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi:10.18653/v1/P18-2006. URL https://aclanthology.org/P18-2006.

Hugh J. Foley, David V. Cross, and Jennifer A. O'reilly. Pervasiveness and magnitude of context effects: Evidence for the relativity of absolute magnitude estimation. *Perception & Psychophysics*, 48(6):551–558, November 1990. ISSN 0031-5117, 1532-5962. doi:10.3758/BF03211601. URL http://link.springer.com/10.3758/BF03211601.

Tal Golan, Prashant C. Raju, and Nikolaus Kriegeskorte. Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proceedings of the National Academy of Sciences*, 117(47):29330–29337, November 2020. ISSN 0027-8424, 1091-6490. doi:10.1073/pnas.1912334117. URL https://www.pnas.org/content/117/47/29330.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Roi Reichart, Sasha Devore, Adeen Flinker, Liat Hasenfratz, Omer Levy, Avinatan Hassidim, Michael Brenner, Yossi Matias, Kenneth A. Norman, Orrin Devinsky, and Uri Hasson. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380, March 2022. ISSN 1097-6256, 1546-1726. doi:10.1038/s41593-022-01026-4. URL https://www.nature.com/articles/s41593-022-01026-4.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6572.

Noah D Goodman and Michael C Frank. Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, 20(11):818–829, 2016.

Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2021. URL https://www.gurobi.com.

Micha Heilbron, Kristijan Armeni, Jan-Mathijs Schoffelen, Peter Hagoort, and Floris P. de Lange. A hierarchy of linguistic predictions during natural language comprehension. preprint, Neuroscience, December 2020. URL http://biorxiv.org/lookup/doi/10.1101/2020.12.03.410399.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Steve R Howell, Damian Jankowicz, and Suzanna Becker. A model of grounded language acquisition: Sensorimotor features improve lexical and grammatical learning. *Journal of Memory and Language*, 53(2):258–276, 2005.

Ann Irvine, Joshua Langfus, and Chris Callison-Burch. The American local news corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1305–1308, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/914_Paper.pdf.

Shailee Jain, Vy Vo, Shivangi Mahto, Amanda LeBel, Javier S Turek, and Alexander Huth. Interpretable multi-timescale models for predicting fmri responses to continuous natural speech. *Advances in Neural Information Processing Systems*, 33:13738–13749, 2020.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online, June 2021. Association for Computational Linguistics. doi:10.18653/v1/2021.naacl-main.324. URL https://aclanthology.org/2021.naacl-main.324.

Reinhard Kneser and Hermann Ney. Improved backing-off for m-gram language modeling. In *1995 international conference on acoustics, speech, and signal processing*, volume 1, pages 181–184. IEEE, 1995.

Bin Liang, Hongcheng Li, Miaoqiang Su, Pan Bian, Xirong Li, and Wenchang Shi. Deep text classification can be fooled. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4208–4215. International Joint Conferences on Artificial Intelligence Organization, 7 2018. doi:10.24963/ijcai.2018/585. URL https://doi.org/10.24963/ijcai.2018/585.

D. V. Lindley. On a Measure of the Information Provided by an Experiment. *The Annals of Mathematical Statistics*, 27 (4):986–1005, 1956. doi:10.1214/aoms/1177728069. URL https://doi.org/10.1214/aoms/1177728069.

Grace W Lindsay. Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of cognitive neuroscience*, 33(10):2017–2031, 2021.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Bingjiang Lyu, William D. Marslen-Wilson, Yuxing Fang, and Lorraine K. Tyler. Finding structure in time: Humans, machines, and language. *bioRxiv*, 2021. doi:10.1101/2021.10.25.465687. URL https://www.biorxiv.org/content/early/2021/12/07/2021.10.25.465687.

Danny Merkx and Stefan L. Frank. Human sentence processing: Recurrence or attention? *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 2021. doi:10.18653/v1/2021.cmcl-1.2. URL http://dx.doi.org/10.18653/v1/2021.cmcl-1.2.

James A. Michaelov, Megan D. Bardolph, Seana Coulson, and Benjamin K. Bergen. Different kinds of cognitive plausibility: why are transformers better than rnns at predicting n400 amplitude? In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43, 2021. URL https://escholarship.org/uc/item/9z06m20f.

John Morris, Eli Lifland, Jack Lanchantin, Yangfeng Ji, and Yanjun Qi. Reevaluating adversarial examples in natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3829–3839, Online, November 2020. Association for Computational Linguistics. doi:10.18653/v1/2020.findings-emnlp.341. URL https://aclanthology.org/2020.findings-emnlp.341.

Nikita Nangia and Samuel R. Bowman. Human vs. muppet: A conservative estimate of human performance on the GLUE benchmark. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4566–4575, Florence, Italy, July 2019. Association for Computational Linguistics. doi:10.18653/v1/P19-1449. URL https://aclanthology.org/P19-1449.

Hamed Nili, Cai Wingfield, Alexander Walther, Li Su, William Marslen-Wilson, and Nikolaus Kriegeskorte. A Toolbox for Representational Similarity Analysis. *PLOS Computational Biology*, 10(4):1–11, 2014. doi:10.1371/journal.pcbi.1003553. URL https://doi.org/10.1371/journal.pcbi.1003553.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models, 2022. URL https://arxiv.org/abs/2202.03286.

Frederike H. Petzschner, Stefan Glasauer, and Klaas E. Stephan. A Bayesian perspective on magnitude estimation. *Trends in Cognitive Sciences*, 19(5):285–293, May 2015. ISSN 13646613. doi:10.1016/j.tics.2015.03.002. URL https://linkinghub.elsevier.com/retrieve/pii/S1364661315000509.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Lara I Rakocevic. *Synthesizing controversial sentences for testing the brain-predictivity of language models*. PhD thesis, Massachusetts Institute of Technology, 2021.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1085–1097, Florence, Italy, July 2019. Association for Computational Linguistics. doi:10.18653/v1/P19-1103. URL https://aclanthology.org/P19-1103.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Semantically equivalent adversarial rules for debugging nlp models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45), 2021.

Heiko H. Schütt, Alexander D. Kipnis, Jörn Diedrichsen, and Nikolaus Kriegeskorte. Statistical inference on representational geometries, 2021. URL https://arxiv.org/abs/2112.09200.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv:1312.6199 [cs]*, December 2013. URL http://arxiv.org/abs/1312.6199. arXiv: 1312.6199.

Mariya Toneva and Leila Wehbe. 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, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/749a8e6c231831ef7756db230b4359c8-Paper.pdf.

Walter JB Van Heuven, Pawel Mandera, Emmanuel Keuleers, and Marc Brysbaert. Subtlex-uk: A new and improved word frequency database for british english. *Quarterly journal of experimental psychology*, 67(6):1176–1190, 2014.

Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. 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, 07 2019. ISSN 2307-387X. doi:10.1162/tacl_a_00279. URL https://doi.org/10.1162/tacl_a_00279.

Alex Wang and Kyunghyun Cho. BERT has a mouth, and it must speak: BERT as a Markov random field language model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 30–36, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi:10.18653/v1/W19-2304. URL https://aclanthology.org/W19-2304.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019a. URL https://proceedings.neurips.cc/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019b. URL https://openreview.net/forum?id=rJ4km2R5t7.

Zhou Wang and Eero P. Simoncelli. Maximum differentiation (MAD) competition: A methodology for comparing computational models of perceptual quantities. *Journal of Vision*, 8(12):8–8, 2008. ISSN 1534-7362. doi:10.1167/8.12.8. URL https://doi.org/10.1167/8.12.8.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392, 07 2020. ISSN 2307-387X. doi:10.1162/tacl_a_00321. URL https://doi.org/10.1162/tacl_a_00321.

Leila Wehbe, Ashish Vaswani, Kevin Knight, and Tom Mitchell. Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 233–243, 2014.

Ethan Wilcox, Pranali Vani, and Roger Levy. A targeted assessment of incremental processing in neural language models and humans. In *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)*, pages 939–952, Online, August 2021. Association for Computational Linguistics. doi:10.18653/v1/2021.acl-long.76. URL https://aclanthology.org/2021.acl-long.76.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Max A. Woodbury. Rank correlation when there are equal variates. *The Annals of Mathematical Statistics*, 11(3): 358–362, 1940. ISSN 00034851. URL http://www.jstor.org/stable/2235684.

Takateru Yamakoshi, Thomas L. Griffiths, and Robert D. Hawkins. Probing BERT's priors with serial reproduction chains, 2022. URL https://arxiv.org/abs/2202.12226.

Wei Emma Zhang, Quan Z Sheng, Ahoud Alhazmi, and Chenliang Li. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(3):1–41, 2020.

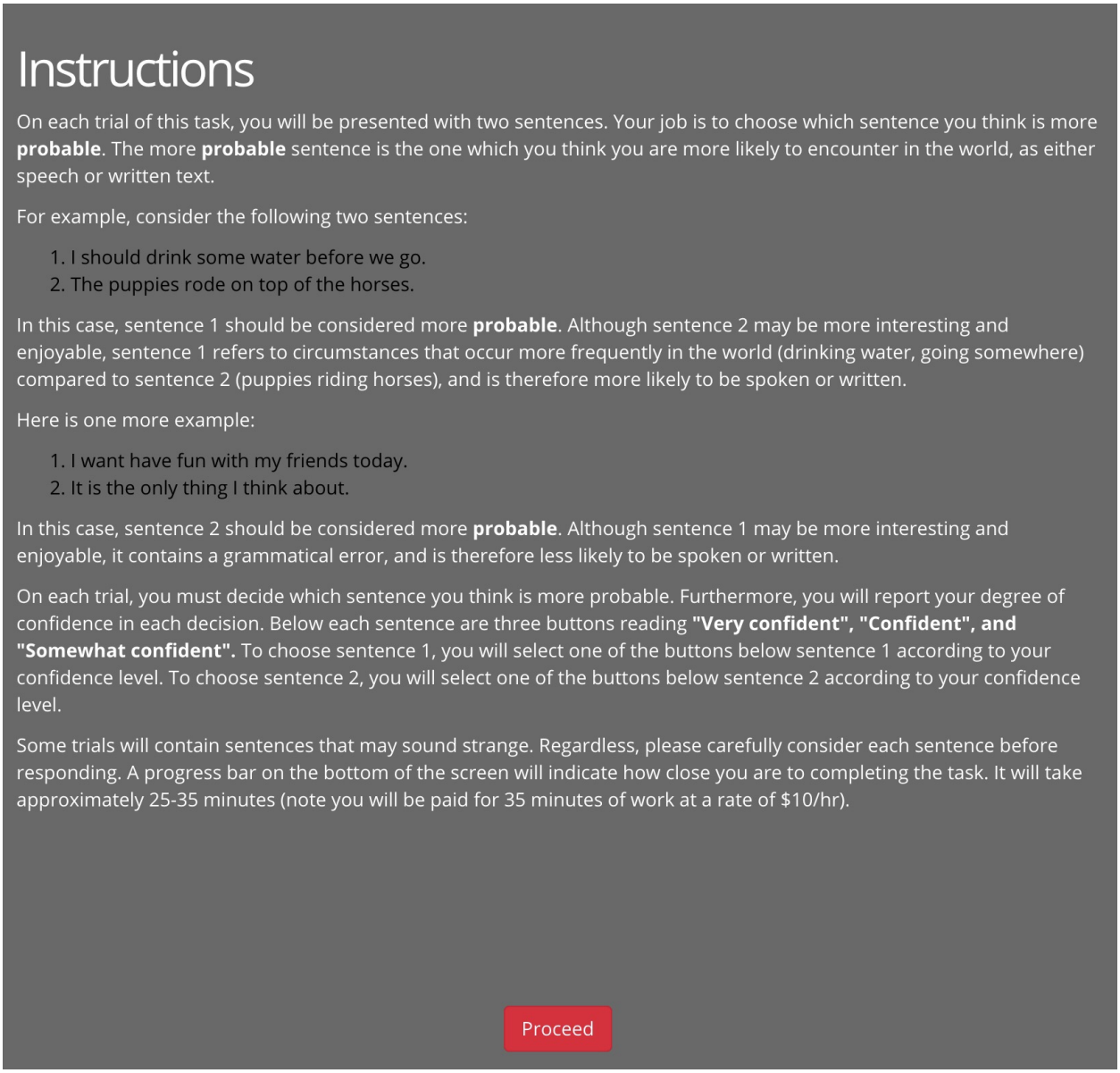


Figure S1: The task instructions provided to the participants at the beginning of the experimental session.

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.

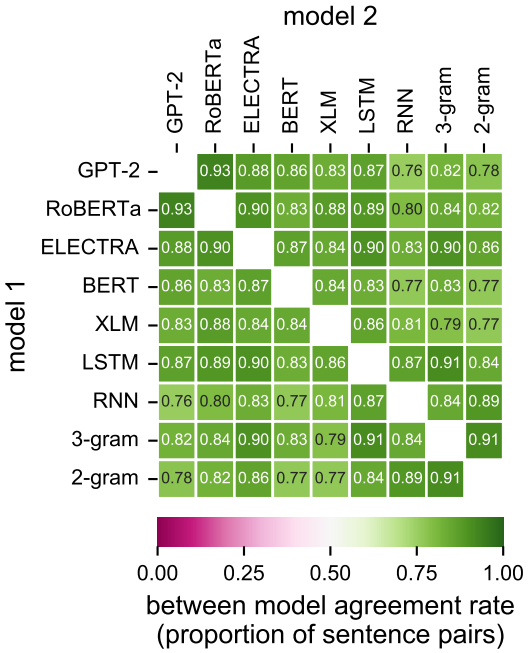
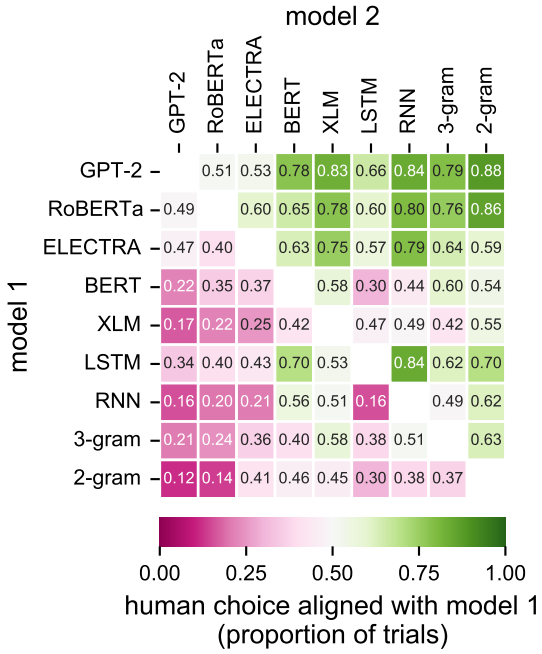


Figure S3: **Between-model agreement rate on the probability ranking of the 90 randomly sampled and paired natural sentence pairs evaluated in the experiment**. Each cell represents the proportion of sentence pairs for which two models make congruent probability ranking (i.e., both models assign a higher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

(a) natural controversial sentences



(b) synthetic controversial sentences

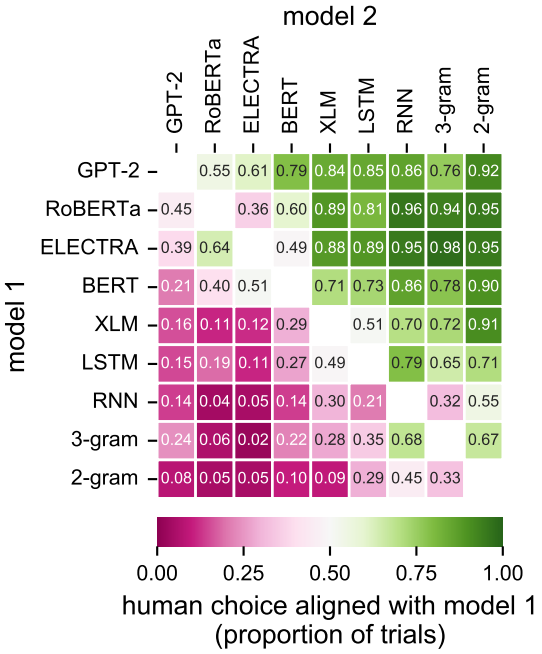


Figure S4: **Pairwise model comparison of model-human consistency.** For each pair of models (represented as one cell in the matrices above), the only trials considered were those in which the stimuli were either selected (a) or synthesized (b) to contrast the predictions of the two models. For these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

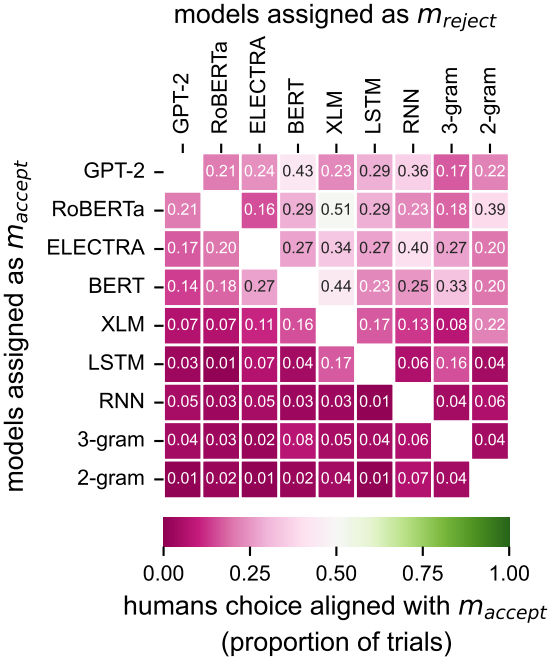


Figure S5: **Pairwise model analysis of human response for natural vs. synthetic sentence pairs.** In each optimization condition, a synthetic sentence s was formed by modifying a natural sentence n so the synthetic sentence would be "rejected" by one model (m_{reject}, columns), minimizing $p(s \mid m_{reject})$, 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.

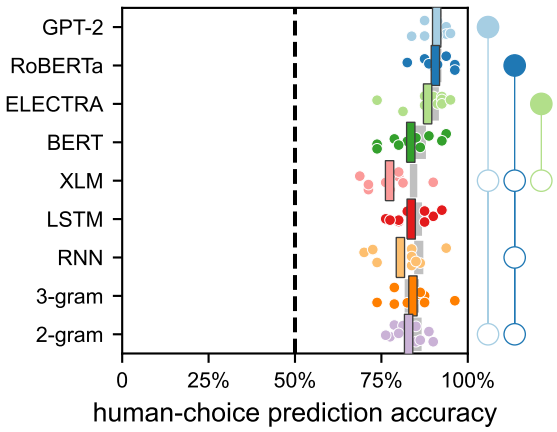


Figure S6: **Model prediction accuracy for pairs of natural and synthetic sentences, evaluating each model across all of the sentence pairs in which it was targeted to rate the synthetic sentence to be less probable than the natural sentence**. The data binning applied here is complementary to the one used in Fig. 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. 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.

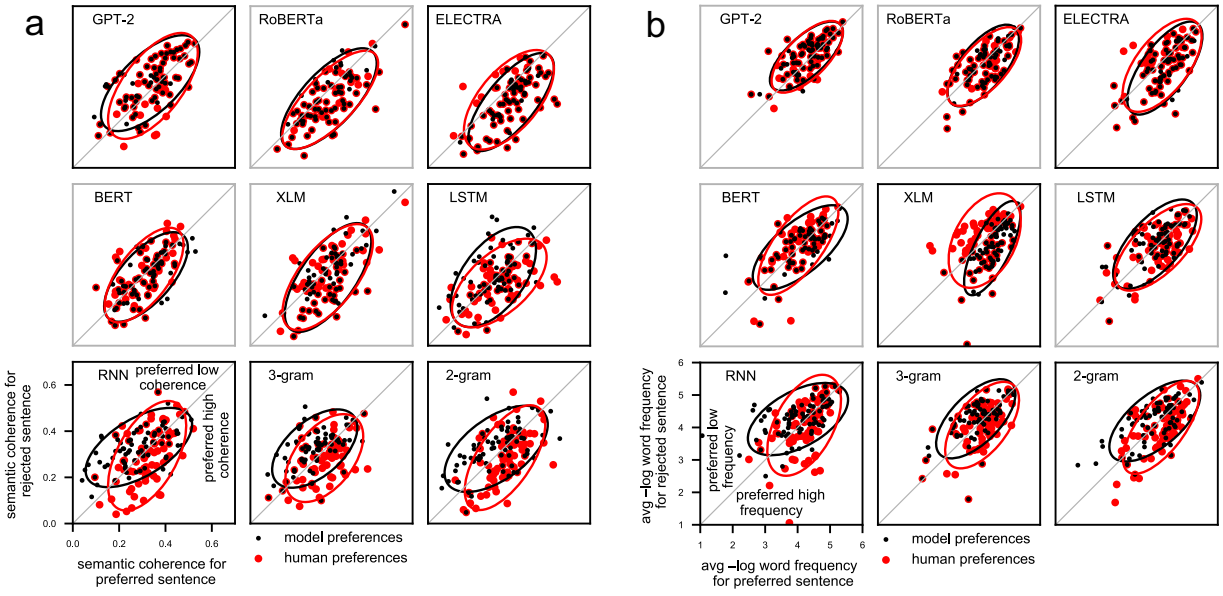


Figure S7: **Linguistic feature values for synthetic sentence pairs. (a)** Semantic coherence values of the preferred and rejected sentence for each synthetic sentence pair. Each panel depicts preferences for both humans (red) and a specific model (black), for sentence pairs that this model was involved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