nature machine intelligence

Article

https://doi.org/10.1038/s42256-023-00718-1

Testing the limits of natural language models for predicting human language judgements

Received: 2 June 2022

Accepted: 11 August 2023

Published online: 14 September 2023

Check for updates

Tal Golan[1,2,6] ✉, Matthew Siegelman[3,6], Nikolaus Kriegeskorte[1,3,4,5] & Christopher Baldassano[3]

Neural network language models appear to be increasingly aligned with how humans process and generate language, but identifying their weaknesses through adversarial examples is challenging due to the discrete nature of language and the complexity of human language perception. We bypass these limitations by turning the models against each other. We generate controversial sentence pairs where two language models disagree about which sentence is more likely to occur. Considering nine language models (including *n*-gram, recurrent neural networks and transformers), we created hundreds of controversial sentence pairs through synthetic optimization or by selecting sentences from a corpus. Controversial sentence pairs proved highly effective at revealing model failures and identifying models that aligned most closely with human judgements of which sentence is more likely. The most human-consistent model tested was GPT-2, although experiments also revealed substantial shortcomings in its alignment with human perception.

Neural network language models are not only key tools in natural language processing (NLP), but are also drawing an increasing scientific interest as potential models of human language processing. Ranging from recurrent neural networks (RNNs)[1,2] to transformers[3–7], each of these language models (explicitly or implicitly) defines a probability distribution over strings of words, predicting which sequences are likely to occur in natural language. There is substantial evidence from measures such as reading times[8], functional magnetic resonance imaging (fMRI)[9], scalp electroencephalograms[10] and intracranial electrocorticography (ECoG)[11] that humans are sensitive to the relative probabilities of words and sentences as captured by language models, even among sentences that are grammatically correct and semantically meaningful. Furthermore, model-derived sentence probabilities can also predict human-graded acceptability judgements[12,13]. These successes, however, have not yet addressed two central questions of interest: (1) which of the models is best aligned with human language processing and (2) how close is the best-aligned model to the goal of fully capturing human judgements?

A dominant approach for evaluating language models is to use a set of standardized benchmarks such as those in the General Language Understanding Evaluation (GLUE)[14], or its successor, SuperGLUE[15]. Though instrumental in evaluating the utility of language models for downstream NLP tasks, these benchmarks prove insufficient for comparing such models as candidate explanations of human language processing. Many components of these benchmarks do not aim to measure human alignment, but instead assess the usefulness of the models' language representation when tuned to a specific downstream task. Some benchmarks challenge language models more directly by comparing the probabilities they assign to grammatical and ungrammatical sentences (for example, BLiMP[16]). However, because such benchmarks are driven by theoretical linguistic considerations, they might fail to detect novel and unexpected ways in which language models may diverge from human language understanding. Finally, an additional practical concern is that the rapid pace of NLP research has led to quick saturation of these types of static benchmark, making it difficult to distinguish between models[17].

[1]Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY, USA. [2]Department of Cognitive and Brain Sciences, Ben-Gurion University of the Negev, Be'er-Sheva, Israel. [3]Department of Psychology, Columbia University, New York, NY, USA. [4]Department of Neuroscience, Columbia University, New York, NY, USA. [5]Department of Electrical Engineering, Columbia University, New York, NY, USA. [6]These authors contributed equally: Tal Golan, Matthew Siegelman. ✉e-mail: golan.neuro@bgu.ac.il

One proposed solution to these issues is the use of dynamic human-in-the-loop benchmarks where people actively stress-test models with an evolving set of tests. However, this approach faces the major obstacle that 'finding interesting examples is rapidly becoming a less trivial task'[17]. We propose to complement human-curated benchmarks with model-driven evaluation. Guided by model predictions rather than experimenter intuitions, we would like to identify particularly informative test sentences, where different models make divergent predictions. This approach of running experiments that are mathematically optimized to 'put in jeopardy' particular models belongs to a long-standing scientific philosophy of design optimization[18]. We can find these critical sentences in large corpora of natural language or synthesize novel test sentences that reveal how different models generalize beyond their training distributions.

In this Article we propose a systematic, model-driven approach for comparing language models in terms of their consistency with human judgements. 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 judgements 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[19]. That work relied on absolute judgements 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[20–22], which have been shown to impact judgements like acceptability[23]. A binary forced-choice behavioural task presenting the participants with a choice between two sentences in each trial, the approach we used here, minimizes the role of between-trial context effects by setting an explicit local context within each trial. Such an approach has been used previously for measuring sentence acceptability[24] and provides substantially more statistical power compared to designs in which acceptability ratings are provided for single sentences[25].

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

Results

We acquired judgements 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 three-point scale (Extended Data Fig. 1 provides a trial example). The experiment was designed to compare nine different language models (Supplementary section 1.1): probability models based on corpus frequencies of two-word and three-word sequences (2-grams and 3-grams) and a range of neural network models including an RNN, a long short-term memory network (LSTM) and five transformer models (BERT, RoBERTa, XLM, ELECTRA and GPT-2).

Efficient model comparison using natural controversial pairs

As a baseline, we randomly sampled and paired eight-word sentences from a corpus of Reddit comments. However, as shown in Fig. 1a, these sentences fail to uncover meaningful differences between the models. For each sentence pair, all models tend to prefer the same sentence (Extended Data Fig. 2), and therefore perform similarly in predicting human preference ratings (Supplementary section 2.1).

We can instead use an optimization procedure (Supplementary section 1.2) to search for controversial sentence pairs, in which one language model assigns a high probability (above the median probability for natural sentences) only to sentence 1 and a second language model assigns a high probability only to sentence 2 (examples are presented in Table 1). Measuring each model's accuracy in predicting human choices for sentence pairs in which it was one of the two targeted models indicated many significant differences in terms of model–human alignment (Fig. 1b), with GPT-2 and RoBERTa showing the best human consistency and 2-gram the worst. We can also compare each model pair separately (using only the stimuli targeting that model pair), yielding a similar pattern of pairwise dominance (Extended Data Fig. 3a). All models except GPT-2, RoBERTa and ELECTRA performed significantly below our lower bound on the noise ceiling (the accuracy obtained by predicting each participant's responses from the other participants' responses), indicating a misalignment between these models' predictions and human judgements that was only revealed when using controversial sentence pairs.

Greater model disentanglement with synthetic 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 natural sentence to drive different models to make opposing predictions, forming synthetic controversial sentences that may lie outside any natural language corpora, as illustrated in Fig. 2 (see Methods, 'Generating synthetic controversial sentence pairs' for full details). Examples of controversial synthetic-sentence pairs that maximally contributed to the models' prediction error are shown in Table 2.

We evaluated how well each model predicted the human sentence choices in all of the controversial synthetic-sentence pairs in which the model was one of the two models targeted (Fig. 3a). This evaluation of model–human alignment resulted in an even greater separation between the models' prediction accuracies than was obtained when using controversial natural-sentence pairs, pushing the weaker models (RNN, 3-gram and 2-gram) far below the 50% chance accuracy level. GPT-2, RoBERTa and ELECTRA were found to be significantly more accurate than the alternative models (BERT, XLM, LSTM, RNN, 3-gram and 2-gram) in predicting the human responses to these trials (with similar results when comparing model pair separately; Extended Data Fig. 3b). All of the models except for GPT-2 were found to be significantly below the lower bound on the noise ceiling, demonstrating misalignment with human judgements.

Pairs of natural and synthetic sentences uncover blindspots

Finally, we considered trials in which the participants were asked to choose between a natural sentence and one of the synthetic sentences, which was generated from that natural sentence. If the language model is fully aligned with human judgements, we would expect humans to agree with the model and select the synthetic sentence at least as much as the natural sentence. In reality, human participants showed a systematic preference for the natural sentences over their synthetic counterparts (Fig. 3b), even when the synthetic sentences were formed such that the stronger models (that is, GPT-2, RoBERTA or ELECTRA) favoured them over the natural sentences (Extended Data Table 1 presents examples). Evaluating natural sentence preference separately for

https://doi.org/10.1038/s42256-023-00718-1

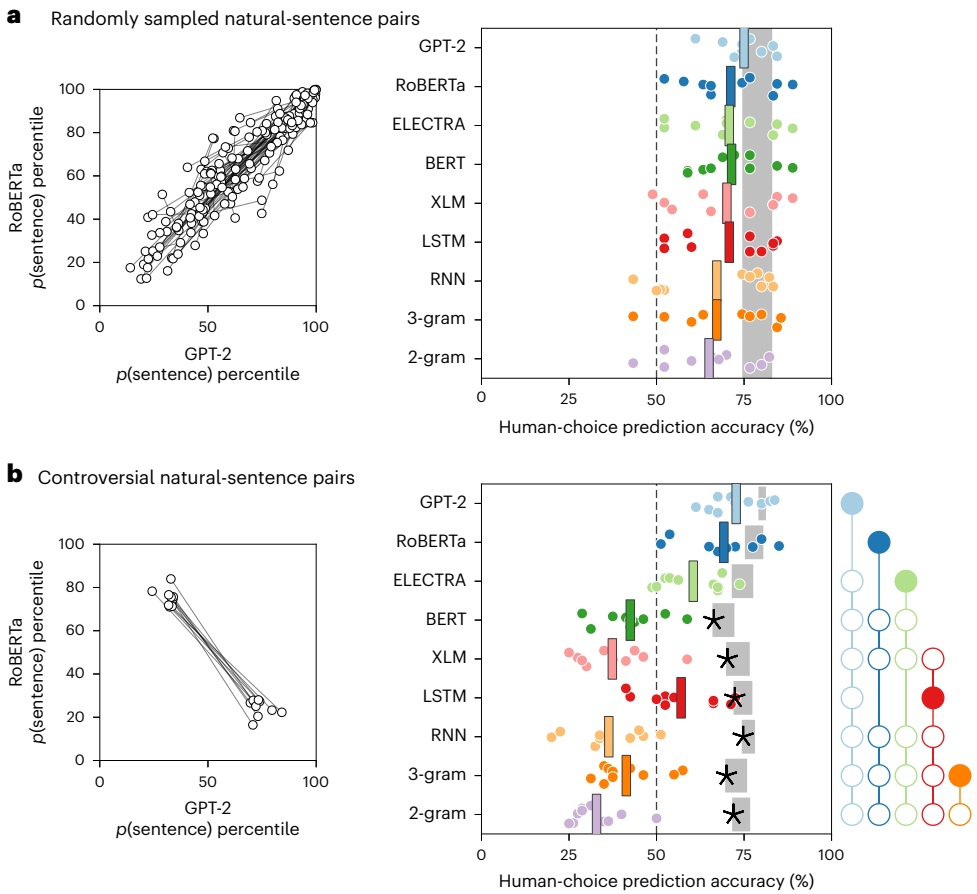


Fig. 1 | Model comparison using natural sentences. a, Left: percentile-transformed sentence probabilities for GPT-2 and RoBERTa (defined relative to all sentences used in the experiment) for randomly sampled pairs of natural sentences. Each pair of connected circles depicts one sentence pair. The two models are highly congruent in their rankings of sentences within a pair (lines have an 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 circle depicts the prediction accuracy of one candidate model averaged across a group of ten participants presented with a unique set of trials. The coloured bars depict grand averages across all 100 participants. The grey bar is the noise ceiling, with its left and right edges being 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 a downward slope). Right: controversial sentence pairs enable efficient model comparison, revealing that BERT, XLM, LSTM, RNN and the *n*-gram models perform significantly below the noise ceiling (asterisks indicate significance—two-sided Wilcoxon signed-rank test, controlling the false discovery rate for nine comparisons at $q < 0.05$). On the right of the plot, each filled circle indicates a model significantly dominating the alternative models, indicated by open circles (two-sided Wilcoxon signed-rank test, controlling the false discovery rate for all 36 model pairs at $q < 0.05$). GPT-2 outperforms all models except RoBERTA at predicting human judgements.

each model pairing (Extended Data Fig. 4), we find that these imperfections can be uncovered even when pairing a strong model with a relatively weak model (such that the strong model accepts the synthetic sentence and the weak model rejects it).

Evaluating the entire dataset reveals a hierarchy of models

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 judgements, here we measure the ordinal correspondence between the graded human choices (taking confidence into account) and the log ratio of the sentence probabilities assigned by each candidate model. Using this more sensitive benchmark (Fig. 4), we found GPT-2 to be the most human-aligned, followed by RoBERTa, then ELECTRA, BERT, XLM and LSTM, and the RNN, 3-gram and 2-gram models. However, all of the models (including GPT-2) were found to be significantly less accurate than the lower bound on the noise ceiling.

One possible reason for the poorer performance of the bidirectional transformers (RoBERTa, ELECTRA, BERT and XLM) compared to the unidirectional transformer (GPT-2) is that computing sentence probabilities in these models is complex, and the probability estimator we developed (see Methods, 'Evaluating sentence probabilities in tranformer models') could be non-optimal; indeed, the popular pseudo-log-likelihood (PLL) approach yields slightly higher accuracy for randomly sampled natural-sentence pairs (Extended Data Fig. 5a). Yet, when we directly compared our estimator to PLL by means of generating and administrating new synthetic controversial sentences, our estimator was found to be markedly better aligned to human judgements (Extended Data Fig. 5b and Extended Data Table 2).

Finally, a control analysis employing probability measures normalized by token count revealed that such normalization had a minimal influence on the observed differences among models (Supplementary section 2.2 and Supplementary Fig. 1).

https://doi.org/10.1038/s42256-023-00718-1

Table 1 | Examples of controversial natural-sentence pairs that maximally contributed to each model's prediction error

Sentence	Log probability (model 1)	Log probability (model 2)	No. of human choices
n_1: Rust is generally caused by salt and sand.	log$p(n_1$|GPT-2)=−50.72	log$p(n_1$|ELECTRA)=**−38.54**	**10**
n_2: Where is Vernon Roche when you need him.	log$p(n_2$|GPT-2)=**−32.26**	log$p(n_2$|ELECTRA)=−58.26	0
n_1: Excellent draw and an overall great smoking experience.	log$p(n_1$|RoBERTa)=−67.78	log$p(n_1$|GPT-2)=**−36.76**	**10**
n_2: I should be higher and tied to inflation.	log$p(n_2$|RoBERTa)=**−54.61**	log$p(n_2$|GPT-2)=−50.31	0
n_1: You may try and ask on their forum.	log$p(n_1$|ELECTRA)=−51.44	log$p(n_1$|LSTM)=**−44.24**	**10**
n_2: I love how they look like octopus tentacles.	log$p(n_2$|ELECTRA)=**−35.51**	log$p(n_2$|LSTM)=−66.66	0
n_1: Grow up and quit whining about minor inconveniences.	log$p(n_1$|BERT)=−82.74	log$p(n_1$|GPT-2)=**−35.66**	**10**
n_2: The extra a is the correct Sanskrit pronunciation.	log$p(n_2$|BERT)=**−51.06**	log$p(n_2$|GPT-2)=−51.10	0
n_1: I like my password manager for this reason.	log$p(n_1$|XLM)=−68.93	log$p(n_1$|RoBERTa)=**−49.61**	**10**
n_2: Kind of like clan of the cave bear.	log$p(n_2$|XLM)=**−44.24**	log$p(n_2$|RoBERTa)=−67.00	0
n_1: We have raised a generation of computer geeks.	log$p(n_1$|LSTM)=−66.41	log$p(n_1$|ELECTRA)=**−36.57**	**10**
n_2: I mean when the refs are being sketchy.	log$p(n_2$|LSTM)=**−42.04**	log$p(n_2$|ELECTRA)=−52.28	0
n_1: This is getting ridiculous and ruining the hobby.	log$p(n_1$|RNN)=−100.65	log$p(n_1$|LSTM)=**−43.50**	**10**
n_2: I think the boys and invincible are better.	log$p(n_2$|RNN)=**−45.16**	log$p(n_2$|LSTM)=−59.00	0
n_1: Then attach them with the supplied wood screws.	log$p(n_1$|3-gram)=−119.09	log$p(n_1$|GPT-2)=**−34.84**	**10**
n_2: Sounds like you were used both a dog.	log$p(n_2$|3-gram)=**−92.07**	log$p(n_2$|GPT-2)=−52.84	0
n_1: Cream cheese with ham and onions on crackers.	log$p(n_1$|2-gram)=−131.99	log$p(n_1$|RoBERTa)=**−54.62**	**10**
n_2: I may have to parallel process that drinking.	log$p(n_2$|2-gram)=**−109.46**	log$p(n_2$|RoBERTa)=−70.69	0

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 in bold), while the model it was pitted against ('model 2') and all ten 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.)

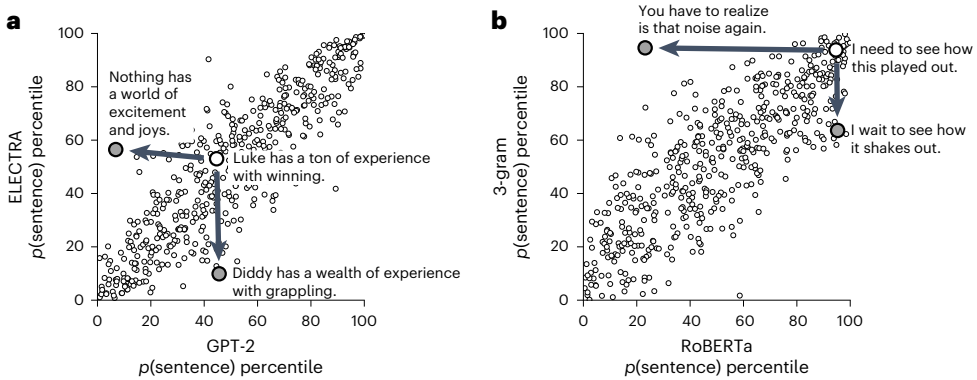


Fig. 2 | Synthesizing controversial sentence pairs. The small open circles denote 500 randomly sampled natural sentences. The big open circle denotes the natural sentence used for initializing the controversial sentence optimization, and the filled circles are the resulting synthetic sentences. **a**, 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 that of GPT-2. **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'.

Discussion

In this study we have probed the ability of language models to predict human relative sentence probability judgements using controversial sentence pairs, selected or synthesized so that two models disagree about which sentence was more probable. We found that (1) GPT-2 (a unidirectional transformer model trained on predicting upcoming tokens) and RoBERTa (a bidirectional transformer trained on a held-out token prediction task) were the most predictive of human judgements 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 judgements on pairs of sentences synthesized to maximize controversiality (Fig. 3a); and (3) GPT-2 was the most human-consistent model when considering the entire behavioural dataset we collected (Fig. 4). However, all of the models, including GPT-2, exhibited behaviour inconsistent with human judgements—using an alternative model as a counterforce, we could

Table 2 | Examples of controversial synthetic-sentence pairs that maximally contributed to each model's prediction error

Sentence	Log probability (model 1)	Log probability (model 2)	No. of human choices
s_1: You can reach his stories on an instant.	log$p$(s_1|GPT-2)=−64.92	log$p$(s_1|RoBERTa)=**−59.98**	**10**
s_2: Anybody can behead a rattles an an antelope.	log$p$(s_2|GPT-2)=**−40.45**	log$p$(s_2|RoBERTa)=−90.87	0
s_1: However they will still compare you to others.	log$p$(s_1|RoBERTa)=−53.40	log$p$(s_1|GPT-2)=**−31.59**	**10**
s_2: Why people who only give themselves to others.	log$p$(s_2|RoBERTa)=**−48.66**	log$p$(s_2|GPT-2)=−47.13	0
s_1: He healed faster than any professional sports player.	log$p$(s_1|ELECTRA)=−48.77	log$p$(s_1|BERT)=**−50.21**	**10**
s_2: One gets less than a single soccer team.	log$p$(s_2|ELECTRA)=**−38.25**	log$p$(s_2|BERT)=−59.09	0
s_1: That is the narrative we have been sold.	log$p$(s_1|BERT)=−56.14	log$p$(s_1|GPT-2)=**−26.31**	**10**
s_2: This is the week you have been dying.	log$p$(s_2|BERT)=**−50.66**	log$p$(s_2|GPT-2)=−39.50	0
s_1: The resilience is made stronger by early adversity.	log$p$(s_1|XLM)=−62.95	log$p$(s_1|RoBERTa)=**−54.34**	**10**
s_2: Every thing is made alive by infinite Ness.	log$p$(s_2|XLM)=**−42.95**	log$p$(s_2|RoBERTa)=−75.72	0
s_1: President Trump threatens to storm the White House.	log$p$(s_1|LSTM)=−58.78	log$p$(s_1|RoBERTa)=**−41.67**	**10**
s_2: West Surrey refused to form the White House.	log$p$(s_2|LSTM)=**−40.35**	log$p$(s_2|RoBERTa)=−67.32	0
s_1: Las beans taste best with a mustard sauce.	log$p$(s_1|RNN)=−131.62	log$p$(s_1|RoBERTa)=**−60.58**	**10**
s_2: Roughly lanes being alive in a statement ratings.	log$p$(s_2|RNN)=**−49.31**	log$p$(s_2|RoBERTa)=−99.90	0
s_1: You are constantly seeing people play the multi.	log$p$(s_1|3-gram)=−107.16	log$p$(s_1|ELECTRA)=**−44.79**	**10**
s_2: This will probably the happiest contradicts the hypocrite.	log$p$(s_2|3-gram)=**−91.59**	log$p$(s_2|ELECTRA)=−75.83	0
s_1: A buyer can own a genuine product also.	log$p$(s_1|2-gram)=−127.35	log$p$(s_1|ELECTRA)=**−40.21**	**10**
s_2: One versed in circumference of highschool I rambled.	log$p$(s_2|2-gram)=**−113.73**	log$p$(s_2|ELECTRA)=−92.61	0

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 in bold), while the model it was pitted against ('model 2') and all ten 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.)

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 computational psycholinguistic modelling

Unlike convolutional neural networks, whose architectural design principles are roughly inspired by biological vision[26], the design of current neural-network language models is largely uninformed by psycholinguistics and neuroscience. However, 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[11,27–35]. We found that RNNs 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[11,32], as well as with evidence from model-based prediction of human reading speed[33,36] and N400 amplitude[36,37]. Among the transformers, GPT-2, RoBERTa and ELECTRA showed the best performance. These models are trained to optimize only word-level prediction tasks, as opposed to BERT and XLM, which are additionally trained on next-sentence prediction and cross-lingual tasks, respectively (and have the same architecture as RoBERTa). This suggests that local word prediction provides better alignment with human language comprehension.

Despite the agreement between our results and previous work in terms of model ranking, the significant failure of GPT-2 in predicting the human responses to natural versus synthetic controversial pairs (Fig. 3b) demonstrates that GPT-2 does not fully emulate the computations employed in human processing of even short sentences. This outcome is in some ways unsurprising, given that GPT-2 (like all of the other models we considered) is an off-the-shelf machine-learning model that was not designed with human psycholinguistic and physiological details in mind. However, the considerable human inconsistency we observed seems to stand in stark contrast with the recent report of GPT-2 explaining about 100% of the explainable variance in fMRI and ECoG responses to natural sentences[32]. Part of this discrepancy could be explained by the fact that Schrimpf et al.[32] mapped GPT-2 hidden-layer activations to brain data by means of regularized linear regression, which can identify a subspace within GPT-2's language representation that is well-aligned with brain responses even if GPT-2's overall sentence probabilities are not human-like. More importantly, when language models are evaluated with natural language, strong statistical models might capitalize on features in the data that are distinct from, but highly correlated with, features that are meaningful to humans. Therefore, a model that performs well on typical sentences might employ computational mechanisms that are very distinct from the brain's, which will only be revealed by testing the model in a more challenging domain. Note that, even the simplest model we considered—a

https://doi.org/10.1038/s42256-023-00718-1

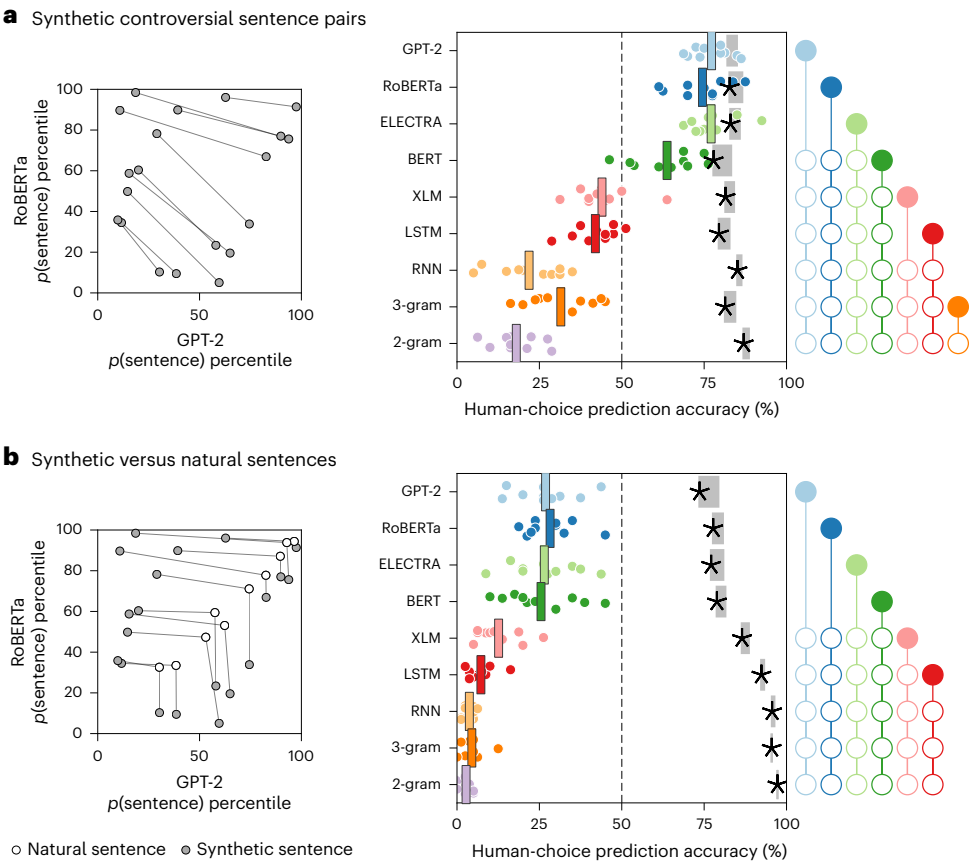


Fig. 3 | Model comparison using synthetic sentences. a, Left: percentile-transformed sentence probabilities for GPT-2 and RoBERTa for controversial synthetic-sentence pairs. Each pair of connected circles depicts one sentence pair. Right: model prediction accuracy, measured as the proportion of trials in which the same sentence was preferred by both the model and the human participant. GPT-2, RoBERTa and ELECTRA significantly outperformed the other models (two-sided Wilcoxon signed-rank test, controlling the false discovery rate for all 36 model comparisons at $q < 0.05$). All of the models except for GPT-2 were found to perform below the noise ceiling (grey) of predicting each participant's choices from the majority votes of the other participants (asterisks indicate significance—two-sided Wilcoxon signed-rank test, controlling the false discovery rate for nine comparisons at $q < 0.05$). **b**, Left: each connected triplet of circles depicts a natural sentence and its derived synthetic sentences, optimized to decrease the probability only under GPT-2 (left circles in a triplet) or only under RoBERTa (bottom circles 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 (Extended Data Fig. 6 presents 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 the caption to Fig. 1 for details on the visualization conventions used in this figure.

2-gram frequency table—actually performed quite well on predicting human judgements for randomly sampled natural sentences, and its deficiencies only became obvious when challenged by controversial sentence pairs. We predict that there will be substantial discrepancies between neural representations and current language models when using stimuli that intentionally stress-test this relationship, using our proposed sentence-level controversiality approach or complementary ideas such as maximizing controversial transition probabilities between consecutive words[38].

Using controversial sentences can be seen as a generalization test of language models: can models predict what kinds of change to a natural sentence will lead to humans rejecting the sentence as improbable? Humans are sometimes capable of comprehending language with atypical constructions (for example, in cases when pragmatic judgements can be made about a speaker's intentions from environmental and linguistic context[39]), but none of the models we tested were fully able to predict which syntactic or semantic perturbations would be accepted or rejected by humans. One possibility is that stronger next-word prediction models, using different architectures, learning rules or training data, might close the gap between models and humans. Alternatively, it might be that optimizing for other linguistic tasks, or even non-linguistic task demands (in particular, representing the external world, the self and other agents), will turn out to be critical for achieving human-like NLP[40].

Controversial sentence pairs as adversarial attacks

Machine vision models are highly susceptible to adversarial examples[41,42]. Such adversarial examples are typically generated by choosing a correctly classified natural image and then searching for a minuscule (and therefore human-imperceptible) image perturbation that would change the targeted model's classification. The prospect that similar covert model failure modes may exist also for language models has motivated proposed generalizations of adversarial methods to textual inputs[43]. However, imperceptible perturbations cannot be applied to written text: any modified word or character is humanly perceptible. Previous work on adversarial examples for language models have instead relied on heuristic constraints aiming to limit the change in the meaning of the text, such as flipping a character[44,45], changing number or gender[46], or replacing words with synonyms[47–49]. However, because these heuristics are only rough approximations of human language

https://doi.org/10.1038/s42256-023-00718-1

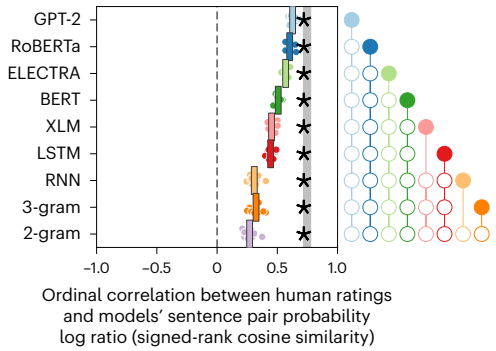


Fig. 4 | Ordinal correlation of the models' sentence probability log ratios and human Likert ratings. For each sentence pair, model prediction was quantified by $\log \frac{p(s^1|m)}{p(s^2|m)}$. This log ratio was correlated with the Likert ratings of each particular participant, using signed-rank cosine similarity (Methods). This analysis, taking all trials and human confidence level into account, indicates that GPT-2 performed best in predicting human sentence probability judgements. However, its predictions are still significantly misaligned with the human choices. See the caption to Fig. 1 for details on the visualization convention.

processing, many of these methods fail to preserve semantic meaning[50]. 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[17,51], but these approaches are inherently slow and costly and are limited by mental models the human subjects form about the evaluated language models.

By contrast, testing language models on controversial sentence pairs does not require approximating or querying a human ground truth during optimization—the objective of controversiality is independent of correctness. Instead, by designing inputs to elicit conflicting predictions among the models and assessing human responses to these inputs only once the optimization loop has terminated, we capitalize on the simple fact that if two models disagree with respect to an input, at least one of the models must be making an incorrect prediction. Pitting language models against other language models can also 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[52]. Our approach shares the underlying principle that an alternative language model can drive a more powerful test than handcrafted heuristics, but here the models have symmetric roles (there are no 'attacking' and 'attacked' models) and we can optimize stimuli directly without relying on filtering.

Limitations and future directions

Although our results demonstrate that using controversial stimuli can identify subtle differences in language models' alignment with human judgements, our study was limited in a number of ways. Our stimuli were all eight-word English sentences, limiting our ability to make cognitively meaningful claims that apply to language use globally. Eight-word sentences are long enough to include common syntactic constructions and convey meaningful ideas, but may not effectively probe long-distance syntactic dependencies[53]. Future work may introduce additional sentence lengths and languages, as well as (potentially adaptive) controversial sentence optimization procedures that consider large sets of candidate models, allowing for greater model coverage than our simpler pairwise approach. Future work may also complement the model-comparative experimental design with procedures designed to identify potential failure modes common to all models.

A more substantial limitation of the current study is that, like any comparison of pre-trained neural networks as potential models of human cognition, there could be multiple reasons (training data, architecture, training tasks and learning rules) why particular models are better aligned with human judgements. For example, as we did not systematically control the training corpora used for training the models, it is possible that some of the observed differences are due to differences in the training sets rather than the model architecture. Therefore, although our results expose failed model predictions, they do not readily answer why these failed predictions arise. Future experiments could compare custom-trained or systematically manipulated models, which reflect specific hypotheses about human language processing. In Extended Data Fig. 5, we demonstrate the power of using synthetic controversial stimuli to conduct sensitive comparisons between models with subtle differences in how sentence probabilities are calculated.

It is important to note that our analyses considered human relative probability judgements as reflecting a scalar measure of acceptability. We made this assumption to bring the language models (which assign a probability measure to each sentence) and the human participants onto a common footing. However, it is possible that different types of sentence pair engage different human cognitive processes. For pairs of synthetic sentences, both sentences may be unacceptable in different ways (for example, exhibit different kinds of grammatical violation), requiring a judgement that weighs the relative importance of multiple dimensions[54] and could therefore produce inconsistent rankings across participants or across trials[55]. By contrast, asking participants to compare a natural sentence and a synthetic sentence (Fig. 3b and Extended Data Table 1) may be more analogous to previous work measuring human acceptability judgements for sentence pairs[24]. Nonetheless, it is worth noting that for all of the controversial conditions, the noise ceiling was significantly above the models' prediction accuracy, indicating non-random human preferences unexplained by current models that should be accounted for by future models, which may have to be more complex and capture multiple processes.

Finally, the use of synthetic controversial sentences can be extended beyond probability judgements. A sufficiently strong language model may enable constraining the experimental design search space to particular sentence distributions (for example, movie reviews or medical questions). Given such a constrained space, we may be able to search for well-formed sentences that elicit contradictory predictions in alternative domain-specific models (for example, sentiment classifiers or question-answering models). However, as indicated by our results, the task of capturing distributions of well-formed sentences is less trivial than it seems.

Methods

Language models

We tested nine models from three distinct classes: *n*-gram models, RNNs and transformers. The *n*-gram models were trained with open-source code from the Natural Language Toolkit[56], the RNNs were trained with architectures and optimization procedures available in PyTorch[57], and the transformers were implemented with the open-source repository HuggingFace[58]. For full details see Supplementary section 1.1.

Evaluating sentence probabilities in transformer models

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

For the set of unidirectional models, evaluating sentence probabilities was performed simply by summing the log probabilities of each successive token in the sentence from left to right, given all the

previous tokens. For bidirectional models, this process was not as straightforward. One challenge is that transformer model probabilities do not necessarily reflect a coherent joint probability; the summed log sentence probability resulting from adding words in one order (for example, left to right) does not necessarily equal the probability resulting from a different order (for example, right to left). Here we developed a novel formulation of bidirectional sentence probabilities in which we considered all permutations of serial word positions as possible construction orders (analogous to the random word visitation order used to sample serial reproduction chains[59]). In practice, we observed that the distribution of log probabilities resulting from different permutations tends to centre tightly around a mean value (for example, for RoBERTa evaluated with natural sentences, the average coefficient of variation was ~0.059). Therefore, to efficiently calculate bidirectional sentence probability, we evaluate 100 different random permutations and define the overall sentence log probability as the mean log probability calculated from each permutation. Specifically, we initialized an eight-word sentence with all tokens replaced with the 'mask' token used in place of to-be-predicted words during model training. We selected a random permutation P of positions 1 to 8, and started by computing the probability of the word at the 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, whereas 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:

$$\begin{aligned}&\log p(w = \text{beehive})\\&= 0.5\left(\log p(w_1 = \text{bee}|w_2 = \text{MASK}) + \log p(w_2 = \text{hive}|w_1 = \text{bee})\right)\\&+0.5\left(\log p(w_2 = \text{hive}|w_1 = \text{MASK}) + \log p(w_1 = \text{bee}|w_2 = \text{hive})\right)\end{aligned} \quad (1)$$

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

One more procedure was applied to ensure that all models were computing a probability distribution over sentences with exactly eight words. When evaluating the conditional probability of a masked word in models with word-piece tokenizers, we normalized the model probabilities to ensure that only single words were being considered, rather than splitting the masked tokens into multiple words. At each evaluation step, each token was restricted to come from one of four normalized distributions: (1) single-mask words were restricted to be tokens with appended white space, (2) masks at the beginning of a word were restricted to be tokens with preceding white space (in models with preceding white space, for example BERT), (3) masks at the end of words were restricted to be tokens with trailing white space (in models with trailing white space, for example XLM) and (4) masks in the middle of words were restricted to tokens with no appended white space.

Assessing potential token count effects on sentence probabilities

Note that, because tokenization schemes varied across models, the number of tokens in a sentence could differ for different models. These alternative tokenizations can be conceived of as different factorizations of the modelled language distribution, changing how a sentence's log probability is additively partitioned across the conditional probability chain (but not affecting its overall probability)[60]. Had we attempted to normalize across models by dividing the log probability by the number of tokens, as is often done when aligning model predictions to human acceptability ratings[12,13], our probabilities would have become strongly tokenization-dependent[60]. To empirically confirm that tokenization differences were not driving our results, we statistically compared the token counts of each model's preferred synthetic sentences with the token counts of their non-preferred counterparts. Although we found significant differences for some of the models, there was no systematic association between token count and model sentence preferences (Supplementary Table 1). In particular, lower sentence probabilities were not systematically confounded by higher token counts.

Defining a shared vocabulary

To facilitate the sampling, selection and synthesis of sentences that could be evaluated by all of the candidate models, we defined a shared vocabulary of 29,157 unique words. Defining this vocabulary was necessary to unify the space of possible sentences between the transformer models (which can evaluate any input due to their word-piece tokenizers) and the neural network and n-gram models (which include whole words as tokens), and to ensure we only included words that were sufficiently prevalent in the training corpora for all models. The vocabulary consisted of the words in the SUBTLEX database[61], after removing words that occurred fewer than 300 times in the 300-million-word corpus (see Supplementary section 1.1) used to train the n-gram and RNN models (that is, with frequencies lower than one in a million).

Sampling of natural sentences

Natural sentences were sampled from the same four text sources used to construct the training corpus for the n-gram and RNN models, while ensuring that there was no overlap between training and testing sentences. Sentences were filtered to include only those with eight distinct words and no punctuation aside from periods, exclamation points or question marks at the end of a sentence. Then, all eight-word sentences were further filtered to include only the words included in the shared vocabulary and to exclude those included in a predetermined list of inappropriate words and phrases. To identify controversial pairs of natural sentences, we used integer linear programming to search for sentences that had above-median probability in one model and minimum probability rank in another model (Supplementary section 1.2).

Generating synthetic controversial sentence pairs

For each pair of models, we synthesized 100 sentence triplets. Each triplet was initialized with a natural sentence n (sampled from Reddit). The words in sentence n were iteratively modified to generate a synthetic sentence with reduced probability according to the first model but not according to the second model. This process was repeated to generate another synthetic sentence from n, in which the roles of the two models were reversed. Conceptually, this approach resembles maximum differentiation (MAD) competition[62], introduced to compare models of image quality assessment. Each synthetic sentence was generated as a solution for a constrained minimization problem:

$$\begin{aligned}s^* = &\operatorname*{argmin}_s \log p\left(s|m_{\text{reject}}\right)\\&\text{subject to } \log p\left(s|m_{\text{accept}}\right) \geq \log p\left(n|m_{\text{accept}}\right)\end{aligned} \quad (2)$$

where 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 as m_{reject}, and for the other synthetic sentence the model roles were flipped.

At each optimization iteration, we selected one of the eight words pseudorandomly (so that all eight positions would be sampled N times before any position would be sampled $N+1$ times) and searched the shared vocabulary for the replacement word that would minimize $\log p(s|m_{\text{reject}})$ 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 (that is, words for which no loss-reducing replacement had been found) have failed in a row.

Word-level search for bidirectional models

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

Since bidirectional 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 $\log p(\text{s}\{i\} \leftarrow w|m)$, the log probability of the sentence s with word w assigned at position i, predicting it from $\log p(\text{s}\{i\} = w|m, s\{i\} \leftarrow \text{mask})$, the completion log probability of word w at position i, given the sentence with the ith word masked:

$$\log \hat{p}(s\{i\} \leftarrow w|m) = \beta_1 \log p(s\{i\} = w|m, s\{i\} \leftarrow \text{mask}) + \beta_0 \quad (3)$$

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

Selecting the best triplets from the optimized sentences

Because 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

$$c_{m_1,m_2}(n, s_1, s_2) = \log \frac{p(n|m_1)}{p(s_1|m_1)} + \log \frac{p(n|m_2)}{p(s_2|m_2)} \quad (4)$$

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 ten 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 ten 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 no. IRB-AAAS0252) and were performed in accordance with the approved protocol. All participants provided prior informed consent. 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 $US5.95. The average participant age was 34.08 ± 12.32 years. The subjects were divided into ten 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 versus synthetic pair, where one model served as m_{accept} and the other as m_{reject}; a natural versus 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. Although each sentence triplet produced by the optimization procedure (section '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 Supplementary Fig. 2). 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-min time limit was enforced for the whole experiment. A progress bar at the bottom of the

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

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

Evaluation of model–human consistency

To measure the alignment on each trial between model judgements and human judgements, 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 favoured by the subject, regardless of the reported confidence level. When both the subject and 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 ten participants who viewed the same set of trials. For the lower bound on the noise ceiling, we predicted each subject's choices from a majority vote of the nine other subjects who were presented with the same trials. For the upper bound (that is, the highest possible accuracy attainable on this data sample), we included the subject themselves in this majority vote-based prediction.

As each of the ten participant groups viewed a unique trial set, these groups provided ten independent replications of the experiment. Models were compared to each other and to the lower bound of the noise ceiling by a Wilcoxon signed-rank test using these ten independent accuracy outcomes as paired samples. For each analysis, the false discovery rate across multiple comparisons was controlled by the Benjamini–Hochberg procedure[63].

In Fig. 4, we instead measure model–human consistency in a more continuous way, comparing the sentence probability ratio in a model to the graded Likert ratings provided by humans (Supplementary section 1.3 presents full details).

Selecting trials for model evaluation

All of the randomly sampled natural-sentence pairs (Fig. 1a) were evaluated for each of the candidate models. Controversial sentence pairs, either natural (Fig. 1b) or synthetic (Fig. 3), were included in a model's evaluation set only if they were formed to target that model specifically. The overall summary analysis (Fig. 4) evaluated all models on all available sentence pairs.

Comparison to pseudo-log-likelihood acceptability measures

Wang and Cho[64] proposed an alternative approach for computing sentence probabilities in bidirectional (BERT-like) models, using a pseudo-log-likelihood measure that simply sums the log probability of each token conditioned on all of the other tokens in the sentence. Although this measure does not reflect a true probability distribution[65], it is positively correlated with human acceptability judgements for several bidirectional models[13,66]. To directly compare this existing approach to our novel method for computing probabilities, we again used the method of controversial sentence pairs to identify the approach most aligned with human judgements. For each bidirectional model (BERT, RoBERTa and ELECTRA), we created two copies of the model, each using a different approach for computing sentence probabilities. We synthesized 40 sentence pairs to maximally differentiate between the two copies of each model, with each copy assigning a higher probability to a different sentence in the pair. Subsequently, we tested 30 human participants, presenting each participant with all 120 sentence pairs. Model–human consistency was quantified as in the main experiment.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this Article.

Data availability

The experimental stimuli, detailed behavioural testing results and code for reproducing all analyses and figures are available at github.com/dpmlab/contstimlang (ref. 67).

Code availability

Sentence optimization code is available at github.com/dpmlab/contstimlang (ref. 67).

References

1. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* **323**, 533–536 (1986).
2. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
3. Devlin, J., Chang, M., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (eds Burstein, J. et al.) 4171–4186 (Association for Computational Linguistics, 2019); https://doi.org/10.18653/v1/n19-1423
4. Liu, Y. et al. RoBERTa: a robustly optimized BERT pretraining approach. Preprint at https://arxiv.org/abs/1907.11692 (2019).
5. Conneau, A. & Lample, G. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems* (eds Wallach, H. et al.) Vol. 32 (Curran Associates, 2019); https://proceedings.neurips.cc/paper/2019/file/c04c19c2c2474dbf5f7ac4372c5b9af1-Paper.pdf
6. Clark, K., Luong, M., Le, Q. V. & Manning, C. D. ELECTRA: pre-training text encoders as discriminators rather than generators. In *Proc. 8th International Conference on Learning Representations ICLR 2020* (ICLR, 2020); https://openreview.net/forum?id=r1xMH1BtvB
7. Radford, A. et al. *Language Models are Unsupervised Multitask Learners* (OpenAI, 2019); https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
8. Goodkind, A. & Bicknell, K. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proc. 8th Workshop on Cognitive Modeling and Computational Linguistics, CMCL 2018* 10–18 (Association for Computational Linguistics, 2018); https://doi.org/10.18653/v1/W18-0102
9. Shain, C., Blank, I. A., Schijndel, M., Schuler, W. & Fedorenko, E. fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia* **138**, 107307 (2020).
10. Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J. & Lalor, E. C. Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Curr. Biol.* **28**, 803–809 (2018).
11. Goldstein, A. et al. Shared computational principles for language processing in humans and deep language models. *Nat. Neurosci.* **25**, 369–380 (2022).

12. Lau, J. H., Clark, A. & Lappin, S. Grammaticality, acceptability and probability: a probabilistic view of linguistic knowledge. *Cogn. Sci.* **41**, 1202–1241 (2017).
13. Lau, J. H., Armendariz, C., Lappin, S., Purver, M. & Shu, C. How furiously can colorless green ideas sleep? Sentence acceptability in context. *Trans. Assoc. Comput. Ling.* **8**, 296–310 (2020).
14. Wang, A. et al. GLUE: a multi-task benchmark and analysis platform for natural language understanding. In *Proc. 7th International Conference on Learning Representations, ICLR 2019* (ICLR, 2019); https://openreview.net/forum?id=rJ4km2R5t7
15. Wang, A. et al. SuperGLUE: a stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems* (eds Wallach, H. et al.) 3266–3280 (Curran Associates, 2019); https://proceedings.neurips.cc/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf
16. Warstadt, A. et al. BLiMP: the benchmark of linguistic minimal pairs for English. *Trans. Assoc. Comput. Ling.* **8**, 377–392 (2020).
17. Kiela, D. et al. Dynabench: rethinking benchmarking in NLP. In *Proc. 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 4110–4124 (Association for Computational Linguistics, 2021); https://doi.org/10.18653/v1/2021.naacl-main.324
18. Box, G. E. P. & Hill, W. J. Discrimination among mechanistic models. *Technometrics* **9**, 57–71 (1967).
19. Golan, T., Raju, P. C. & Kriegeskorte, N. Controversial stimuli: pitting neural networks against each other as models of human cognition. *Proc. Natl Acad. Sci. USA* **117**, 29330–29337 (2020).
20. Cross, D. V. Sequential dependencies and regression in psychophysical judgments. *Perception Psychophys.* **14**, 547–552 (1973).
21. Foley, H. J., Cross, D. V. & O'reilly, J. A. Pervasiveness and magnitude of context effects: evidence for the relativity of absolute magnitude estimation. *Perception Psychophys.* **48**, 551–558 (1990).
22. Petzschner, F. H., Glasauer, S. & Stephan, K. E. A Bayesian perspective on magnitude estimation. *Trends Cogn. Sci.* **19**, 285–293 (2015).
23. Greenbaum, S. Contextual influence on acceptability judgments. *Linguistics* **15**, 5–12 (1977).
24. Schütze, C. T. & Sprouse, J. in *Research Methods in Linguistics* (eds Podesva, R. J. & Sharma, D.) 27–50 (Cambridge Univ. Press, 2014); https://doi.org/10.1017/CBO9781139013734.004
25. Sprouse, J. & Almeida, D. Design sensitivity and statistical power in acceptability judgment experiments. *Glossa* **2**, 14 (2017).
26. Lindsay, G. W. Convolutional neural networks as a model of the visual system: past, present and future. *J. Cogn. Neurosci.* **33**, 2017–2031 (2021).
27. Wehbe, L., Vaswani, A., Knight, K. & Mitchell, T. Aligning context-based statistical models of language with brain activity during reading. In *Proc. 2014 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*) 233–243 (Association for Computational Linguistics, 2014); https://doi.org/10.3115/v1/D14-1030
28. Toneva, M. & Wehbe, L. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Advances in Neural Information Processing Systems* (eds Wallach, H. et al.) Vol. 32 (Curran Associates, 2019); https://proceedings.neurips.cc/paper/2019/file/749a8e6c231831ef7756db230b4359c8-Paper.pdf
29. Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P. & De Lange, F. P. A hierarchy of linguistic predictions during natural language comprehension. *Proc. Natl Acad. Sci. USA* **119**, 2201968119 (2022).
30. Jain, S. et al. Interpretable multi-timescale models for predicting fMRI responses to continuous natural speech. In *Advances in Neural Information Processing Systems* (eds Larochelle, H. et al.) Vol. 33, 13738–13749 (Curran Associates, 2020); https://proceedings.neurips.cc/paper_files/paper/2020/file/9e9a30b74c49d07d8150c8c83b1ccf07-Paper.pdf
31. Lyu, B., Marslen-Wilson, W. D., Fang, Y. & Tyler, L. K. Finding structure in time: humans, machines and language. Preprint at https://www.biorxiv.org/content/10.1101/2021.10.25.465687v2 (2021).
32. Schrimpf, M. et al. The neural architecture of language: integrative modeling converges on predictive processing. *Proc. Natl Acad. Sci. USA* **118**, 2105646118 (2021).
33. Wilcox, E., Vani, P. & Levy, R. A targeted assessment of incremental processing in neural language models and humans. In *Proc. 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (*Volume 1*: *Long Papers*) 939–952 (Association for Computational Linguistics, 2021); https://doi.org/10.18653/v1/2021.acl-long.76
34. Caucheteux, C. & King, J.-R. Brains and algorithms partially converge in natural language processing. *Commun. Biol.* **5**, 134 (2022).
35. Arehalli, S., Dillon, B. & Linzen, T. Syntactic surprisal from neural models predicts, but underestimates, human processing difficulty from syntactic ambiguities. In *Proc. 26th Conference on Computational Natural Language Learning* (*CoNLL*) 301–313 (Association for Computational Linguistics, 2022); https://aclanthology.org/2022.conll-1.20
36. Merkx, D. & Frank, S. L. Human sentence processing: recurrence or attention? In *Proc. Workshop on Cognitive Modeling and Computational Linguistics* 12–22 (Association for Computational Linguistics, 2021); https://doi.org/10.18653/v1/2021.cmcl-1.2
37. Michaelov, J. A., Bardolph, M. D., Coulson, S. & Bergen, B. K. Different kinds of cognitive plausibility: why are transformers better than RNNs at predicting N400 amplitude? In *Proc. Annual Meeting of the Cognitive Science Society* Vol. 43 (2021); https://escholarship.org/uc/item/9z06m20f
38. Rakocevic, L. I. Synthesizing controversial sentences for testing the brain-predictivity of language models. PhD thesis, Massachusetts Institute of Technology (2021); https://hdl.handle.net/1721.1/130713
39. Goodman, N. D. & Frank, M. C. Pragmatic language interpretation as probabilistic inference. *Trends Cogn. Sci.* **20**, 818–829 (2016).
40. Howell, S. R., Jankowicz, D. & Becker, S. A model of grounded language acquisition: sensorimotor features improve lexical and grammatical learning. *J. Mem. Lang.* **53**, 258–276 (2005).
41. Szegedy, C. et al. Intriguing properties of neural networks. Preprint at http://arxiv.org/abs/1312.6199 (2013).
42. Goodfellow, I. J., Shlens, J. & Szegedy, C. Explaining and harnessing adversarial examples. In Proc. *3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings* (2015); http://arxiv.org/abs/1412.6572
43. Zhang, W. E., Sheng, Q. Z., Alhazmi, A. & Li, C. Adversarial attacks on deep-learning models in natural language processing: a survey. *ACM Trans. Intell. Syst. Technol.* **11**, 1–41 (2020).
44. Liang, B. et al. Deep text classification can be fooled. In *Proc. Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18* 4208–4215 (International Joint Conferences on Artificial Intelligence Organization, 2018); https://doi.org/10.24963/ijcai.2018/585
45. Ebrahimi, J., Rao, A., Lowd, D. & Dou, D. HotFlip: white-box adversarial examples for text classification. In *Proc. 56th Annual Meeting of the Association for Computational Linguistics* (*Volume 2*: *Short Papers*) 31–36 (Association for Computational Linguistics, 2018); https://doi.org/10.18653/v1/P18-2006

46. Abdou, M. et al. The sensitivity of language models and humans to Winograd schema perturbations. In *Proc. 58th Annual Meeting of the Association for Computational Linguistics* 7590–7604 (Association for Computational Linguistics, 2020); https://doi.org/10.18653/v1/2020.acl-main.679
47. Alzantot, M. et al. Generating natural language adversarial examples. In *Proc. 2018 Conference on Empirical Methods in Natural Language Processing* 2890–2896 (Association for Computational Linguistics, 2018); https://doi.org/10.18653/v1/D18-1316
48. Ribeiro, M. T., Singh, S. & Guestrin, C. Semantically equivalent adversarial rules for debugging NLP models. In *Proc. 56th Annual Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*) 856–865 (Association for Computational Linguistics, 2018); https://doi.org/10.18653/v1/P18-1079
49. Ren, S., Deng, Y., He, K. & Che, W. Generating natural language adversarial examples through probability weighted word saliency. In *Proc. 57th Annual Meeting of the Association for Computational Linguistics* 1085–1097 (Association for Computational Linguistics, 2019); https://doi.org/10.18653/v1/P19-1103
50. Morris, J., Lifland, E., Lanchantin, J., Ji, Y. & Qi, Y. Reevaluating adversarial examples in natural language. In *Findings of the Association for Computational Linguistics*: *EMNLP 2020* 3829–3839 (Association for Computational Linguistics, 2020); https://doi.org/10.18653/v1/2020.findings-emnlp.341
51. Wallace, E., Rodriguez, P., Feng, S., Yamada, I. & Boyd-Graber, J. Trick me if you can: human-in-the-loop generation of adversarial examples for question answering. *Trans. Assoc. Comput. Ling.* **7**, 387–401 (2019).
52. Perez, E. et al. Red teaming language models with language models. In *Proc.of the 2022 Conference on Empirical Methods in Natural Language Processing* 3419–3448 (Association for Computational Linguistics, 2022); https://doi.org/10.18653/v1/2022.emnlp-main.225
53. Gibson, E. Linguistic complexity: locality of syntactic dependencies. *Cognition* **68**, 1–76 (1998).
54. Watt, W. C. The indiscreteness with which impenetrables are penetrated. *Lingua* **37**, 95–128 (1975).
55. Schütze, C. T. *The Empirical Base of Linguistics*, Classics in Linguistics Vol. 2 (Language Science Press, 2016); https://doi.org/10.17169/langsci.b89.100
56. Bird, S., Klein, E. & Loper, E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit* ('O'Reilly Media, 2009).
57. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems* (eds Wallach, H. et al.) Vol. 32, 8024–8035 (Curran Associates, 2019); http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
58. Wolf, T. et al. Transformers: state-of-the-art natural language processing. In *Proc. 2020 Conference on Empirical Methods in Natural Language Processing*: *System Demonstrations* 38–45 (Association for Computational Linguistics, 2020); https://doi.org/10.18653/v1/2020.emnlp-demos.6
59. Yamakoshi, T., Griffiths, T. & Hawkins, R. Probing BERT's priors with serial reproduction chains. In *Findings of the Association for Computational Linguistics*, *ACL 2022* 3977–3992 (Association for Computational Linguistics, 2022); https://doi.org/10.18653/v1/2022.findings-acl.314
60. Chestnut, S. Perplexity https://drive.google.com/uc?export=download&id=1gSNfGQ6LPxlNctMVwUKrQpUA7OLZ83PW (accessed 23 September 2022).
61. Heuven, W. J. B., Mandera, P., Keuleers, E. & Brysbaert, M. Subtlex-UK: a new and improved word frequency database for British English. *Q. J. Exp. Psychol.* **67**, 1176–1190 (2014).
62. Wang, Z. & Simoncelli, E. P. Maximum differentiation (MAD) competition: a methodology for comparing computational models of perceptual quantities. *J. Vision* **8**, 8 (2008).
63. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B (Methodol.)* **57**, 289–300 (1995).
64. Wang, A. & Cho, K. BERT has a mouth, and it must speak: BERT as a Markov random field language model. In *Proc. Workshop on Methods for Optimizing and Evaluating Neural Language Generation* 30–36 (Association for Computational Linguistics, 2019); https://doi.org/10.18653/v1/W19-2304
65. Cho, K. BERT has a mouth and must speak, but it is not an MRF https://kyunghyuncho.me/bert-has-a-mouth-and-must-speak-but-it-is-not-an-mrf/ (accessed 28 September 2022).
66. Salazar, J., Liang, D., Nguyen, T. Q. & Kirchhoff, K. Masked language model scoring. In *Proc. 58th Annual Meeting of the Association for Computational Linguistics* 2699–2712 (Association for Computational Linguistics, 2020); https://doi.org/10.18653/v1/2020.acl-main.240
67. Golan, T., Siegelman, M., Kriegeskorte, N. & Baldassano, C. Code and data for 'Testing the limits of natural language models for predicting human language judgments' (Zenodo, 2023); https://doi.org/10.5281/zenodo.8147166

Acknowledgements

This material is based on work partially supported by the National Science Foundation under grant no. 1948004 to N.K. This publication was made possible with the support of the Charles H. Revson Foundation (to T.G.). The statements made and views expressed, however, are solely the responsibility of the authors.

Author contributions

T.G., M.S., N.K. and C.B. designed the study. M.S. implemented the computational models and T.G. implemented the sentence pair optimization procedures. M.S. conducted the behavioural experiments. T.G. and M.S. analysed the experiments' results. T.G., M.S., N.K. and C.B. wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s42256-023-00718-1.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-023-00718-1.

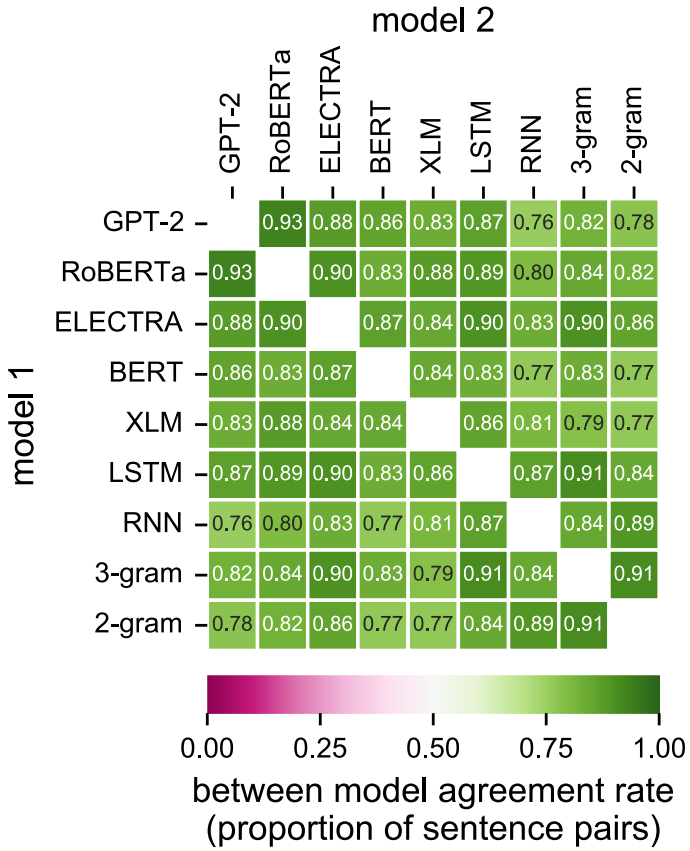
Correspondence and requests for materials should be addressed to Tal Golan.

Peer review information *Nature Machine Intelligence* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editor: Jacob Huth, in collaboration with the *Nature Machine Intelligence* team.

https://doi.org/10.1038/s42256-023-00718-1

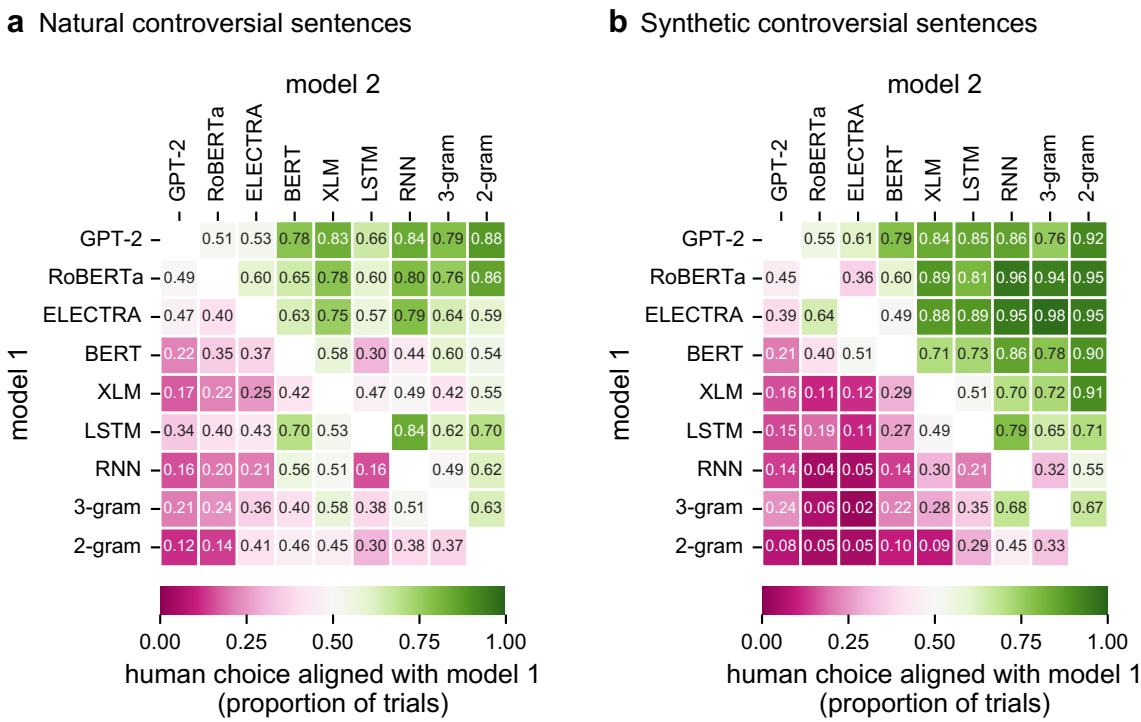
Extended Data Fig. 1 | An example of one experimental trial, as presented to the participants. The participant must choose one sentence while providing their confidence rating on a 3-point scale.

https://doi.org/10.1038/s42256-023-00718-1



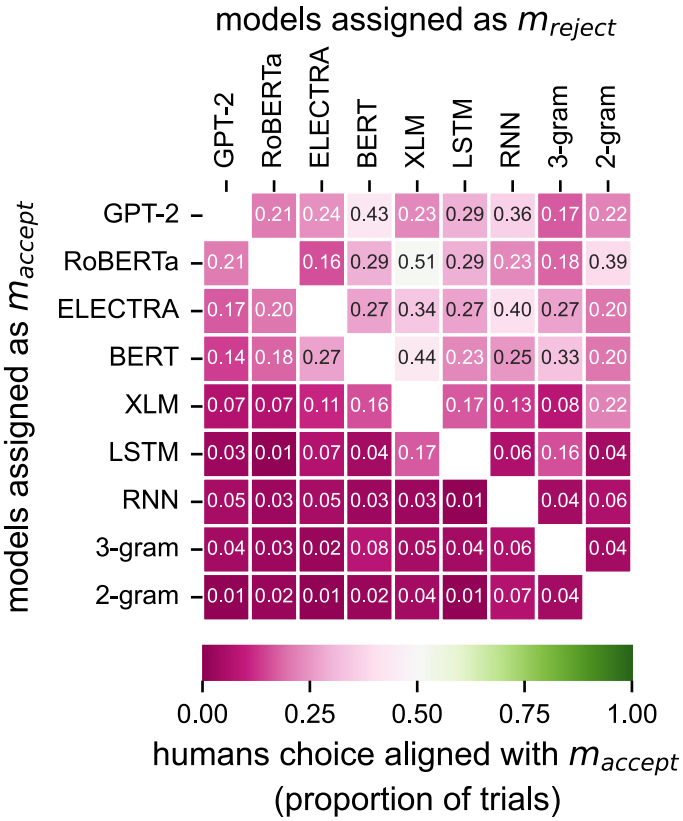
Extended Data Fig. 2 | Between-model agreement rate on the probability ranking of the 90 randomly sampled and paired natural sentence pairs evaluated in the experiment. Each cell represents the proportion of sentence pairs for which two models make congruent probability ranking (that is, both models assign a higher probability to sentence 1, or both models assign a higher probability to sentence 2).

https://doi.org/10.1038/s42256-023-00718-1



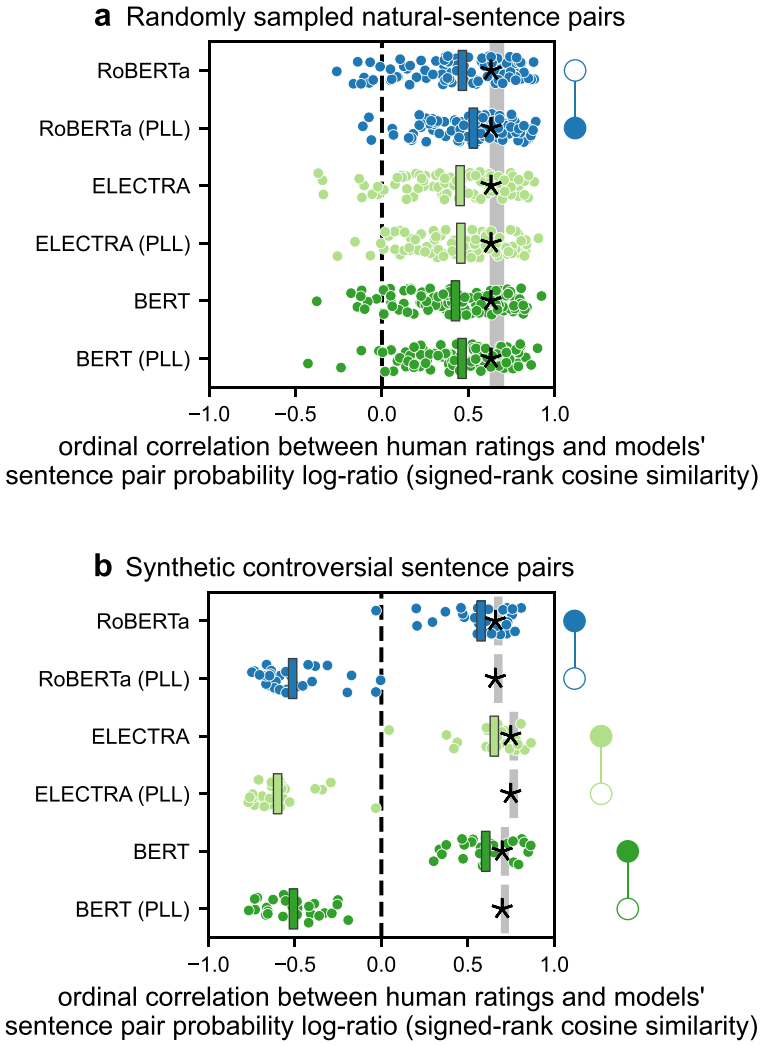
Extended Data Fig. 3 | Pairwise model comparison of model-human consistency. For each pair of models (represented as one cell in the matrices above), the only trials considered were those in which the stimuli were either selected (a) or synthesized (b) to contrast the predictions of the two models. For these trials, the two models always made controversial predictions (that is, 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.

https://doi.org/10.1038/s42256-023-00718-1



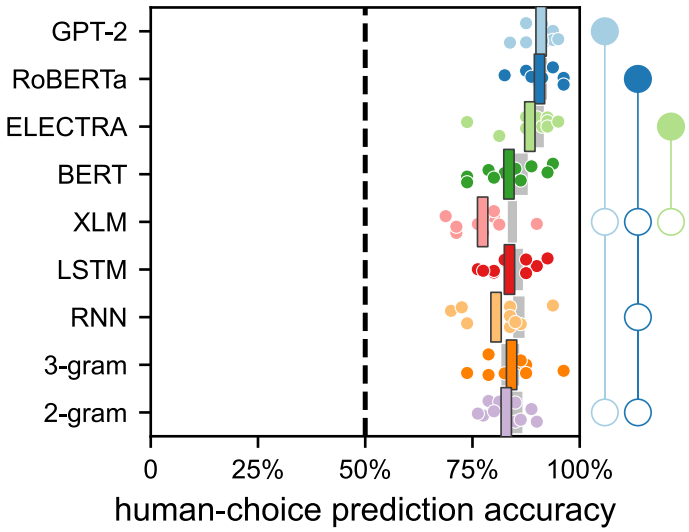
Extended Data Fig. 4 | Pairwise model analysis of human response for natural vs. synthetic sentence pairs. In each optimization condition, a synthetic sentence *s* was formed by modifying a natural sentence *n* so the synthetic sentence would be 'rejected' by one model (m_{reject}, columns), minimizing $p(s|m_{reject})$, and would be 'accepted' by another model (m_{accept}, rows), satisfying the constraint $p(s|m_{accept}) \geq p(n|m_{accept})$. 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.

https://doi.org/10.1038/s42256-023-00718-1



Extended Data Fig. 5 | Human consistency of bidirectional transformers: approximate log-likelihood versus pseudo-log-likelihood (PLL). Each dot in the plots above depicts the ordinal correlation between the judgments of one participant and the predictions of one model. (**a**) The performance of BERT, RoBERTa, and ELECTRA in predicting the human judgments of randomly sampled natural sentence pairs in the main experiment, using two different likelihood measures: our novel approximate likelihood method (that is, averaging multiple conditional probability chains, see Methods) and pseudo-likelihood (PLL, summating the probability of each word given all of the other words[64]). For each model, we statistically compared the two likelihood measures to each other and to the noise ceiling using a two-sided Wilcoxon signed-rank test across the participants. False discovery rate was controlled at $q < 0.05$ for the 9 comparisons. **When predicting human preferences of natural sentences, the pseudo-log-likelihood measure is at least as accurate as our proposed approximate log-likelihood measure. (b)** Results from a follow-up experiment, in which we synthesized synthetic sentence pairs for each of the model pairs, pitting the two alternative likelihood measures against each other. Statistical testing was conducted in the same fashion as in panel a. These results indicate that for each of the three bidirectional language models, the approximate log-likelihood measure is considerably and significantly ($q < 0.05$) more human-consistent than the pseudo-likelihood measure. **Synthetic controversial sentence pairs uncover a dramatic failure mode of the pseudo-log-likelihood measure, which remains covert when the evaluation is limited to randomly-sampled natural sentences**. See Extended Data Table 2 for synthetic sentence pair examples.

https://doi.org/10.1038/s42256-023-00718-1



Extended Data Fig. 6 | 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.

https://doi.org/10.1038/s42256-023-00718-1

Extended Data Table 1 | Examples of pairs of synthetic and natural sentences that maximally contributed to each model's prediction error

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

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.

https://doi.org/10.1038/s42256-023-00718-1

Extended Data Table 2 | Examples of controversial synthetic-sentence pairs that maximally contributed to the prediction error of bidirectional transformers using pseudo-log-likelihood (PLL)

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

For each bidirectional model, the table displays two sentence pairs on which the model failed severely when its prediction was based on pseudo-log-likelihood (PLL) estimates[64]. In each of these sentence pairs, the PLL estimate favors sentence s_2 (higher PLL bolded), while the approximate log-likelihood estimate and most of the human subjects presented with that sentence pair preferred sentence s_1. (When more than one sentence pair induced an equal maximal error in a model, the example included in the table was chosen at random.) **Sentences with long, multi-token words (for example, 'methamphetamine') have high PLL estimates since each of their tokens is well predicted by the other tokens. And yet, the entire sentence is improbable according to human judgments and approximate log-probability estimates based on proper conditional probability chains**.

nature research

Corresponding author(s): Tal Golan

Last updated by author(s): 7/20/2023

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The behavioral data was collected using a custom Gorilla script (https://gorilla.sc/). The Gorilla code will be provided upon request.
Data analysis	Data were analyzed with custom Python code employing the pandas and statsmodels libraries. Our complete analysis code is shared online at https://github.com/dpmlab/contstimlang. The repository also includes the code necessary for generating the controversial sentence pairs.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The experimental stimuli, detailed behavioral testing results, and code for reproducing all analyses and figures are available at https://github.com/dpmlab/contstimlang}{github.com/dpmlab/contstimlang.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Quantitative experimental study.
Research sample	Native English speaking US-based prolific.co users. 55 males and 45 females. The average participant age was 34.08 ± 12.32. We chose to sample US-based native English speakers since this is the reference population for most of the language models we considered. Since the participants were free to decide whether to participate in our study, the sample is not necessarily representative. A follow up experiment (presented in Extended Data Figure 5 and described in detail in supplementary section 1.2) recruited additional 30 participants from the same sampling frame.
Sampling strategy	No formal subject sampling was employed (online requirement continued until the study was completed). The number of subjects was set before data collection according to our experience with similar designs (e.g., Golan, Raju & Kriegeskorte, 2020 PNAS).
Data collection	Data was collected online at the privacy of the participants' homes. The researchers were not present or involved in the experimental sessions (i.e., the behavioral experiment was fully automated).
Timing	June 14 2021 through August 8 2021. The follow up experiment was conducted from October 25 2022 through November 11 2022.
Data exclusions	We used a pre-established exclusion criteria to ensure that all analyzed participants showed sincere effort in their linguistic judgments. We included 12 trials with pairs of a naturally occurring sentences and their shuffled versions. 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. Using the same criteria, we rejected the data of 3 participants in the follow up experiment.
Non-participation	5 participants failed to complete the main experiment.
Randomization	In the main experiment, we randomly allocated participants to replication groups (every ten subjects were presented with a different stimulus set).

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	See above

Recruitment	Participants were recruited through the Prolific.co website. We did not identify a particular self-selection bias that is likely to impact the results.
Ethics oversight	The Columbia University Institutional Review Board (protocol number IRB-AAAS0252).

Note that full information on the approval of the study protocol must also be provided in the manuscript.