Rarely a problem? Language models exhibit inverse scaling in their predictions following *few*-type quantifiers

James A. Michaelov
Department of Cognitive Science
University of California, San Diego
j1michae@ucsd.edu

Benjamin K. Bergen
Department of Cognitive Science
University of California, San Diego
bkbergen@ucsd.edu

Abstract

How well do language models deal with quantification? In this study, we focus on *few*-type quantifiers, as in *few children like toys*, which might pose a particular challenge for language models because the sentence components without the quantifier are likely to co-occur, and *few*-type quantifiers are rare. We present 960 English sentence stimuli from two human neurolinguistic experiments to 22 autoregressive transformer models of differing sizes. Not only do all the models perform poorly on *few*-type quantifiers, but overall the larger the model, the worse its performance. This inverse scaling is consistent with previous work suggesting that larger models increasingly reflect online rather than offline human processing, and we argue that the decreasing performance of larger models may challenge uses of language models as the basis for natural language systems.

1 Introduction

Quantifiers can dramatically alter the meaning of an utterance. Consider the sentences in (1).

(1) (a) Most sharks are harmless.
(b) Most sharks are dangerous.
(c) Few sharks are harmless.
(d) Few sharks are dangerous.

Despite the fact that (a) and (c) have the same content words in the same syntactic arrangement, the statements have starkly different meanings. The same is true of (b) and (d). Being able to successfully comprehend these differences is useful, and in an example such as this one, vitally important[1].

Yet current work suggests that language models deal poorly with quantifiers—they struggle to predict which quantifier is used in a given context (Pezzelle et al., 2018; Talmor et al., 2020), and also perform poorly at generating appropriate continuations following logical quantifiers (Kalouli et al., 2022). This is especially concerning given the recent trend of using large language models (sometimes referred to as 'foundation models'; Bommasani et al., 2021) as general systems that can perform multiple tasks, including question answering, without specific training (Brown et al., 2020; Raffel et al., 2020; Lin et al., 2021; Srivastava et al., 2022; Hoffmann et al., 2022; Rae et al., 2022; Zhang et al., 2022; Chowdhery et al., 2022). It is thus crucial that such systems be able to distinguish among sentences like those in (1) in human-like ways both during training and when generating responses.

The aim of the present study is to evaluate how well language models take into account the meaning of a quantifier when generating the text that follows it, and to investigate whether this scales with model size. We are particularly interested in the question of whether language models exhibit *inverse scaling*—that is, whether as model size increases, performance decreases rather than increases (Perez et al., 2022; McKenzie et al., 2022a). Inverse scaling is an issue of serious concern for developing and training new language models, since inverse scaling could indicate 'outer misalignment' (Perez et al., 2022)—that the training approach is leading to models that produce undesirable outputs, which may get worse as performance at training objectives increases. Inverse scaling is also a concern for models' ultimate use. As models increase in size and perform better at a wider range of benchmarks (for recent examples, see, e.g., Srivastava et al., 2022; Chowdhery et al., 2022), they may be increasingly assumed to be trustworthy and general-purpose, and thus able to perform well tasks on which they have not been tested (Raji et al., 2021). This could lead to a range of possible harms, from misidentifying whether something is dangerous or not (as in the opening example), to amplifying biases (Bender et al., 2021).

[1]Note that most sharks are in fact harmless to humans; see, e.g., https://www.floridamuseum.ufl.edu/discover-fish/sharks/shark-attack-faq/.

Findings of the Association for Computational Linguistics: ACL 2023, pages 14162–14174
July 9-14, 2023

To test how well language models deal with quantifiers, we follow the approach of Ettinger (2020) in using sentences from a study on human language comprehension to inform our evaluation. Ettinger (2020) found that following a negation, the predictions of $\text{BERT}_{\text{BASE}}$ and $\text{BERT}_{\text{LARGE}}$ in simple sentences expressing a proposition with or without negation (from Fischler et al., 1984) do not appear sensitive to negation—for example, $\text{BERT}_{\text{LARGE}}$ predicts the final word of *a robin is a* ***bird*** to be more likely than *a robin is a* ***tree***, but also predicts that *a robin is not a* ***bird*** is more likely than *a robin is not a* ***tree***. In this way, the models' predictions more closely match those made by humans 'online'—that is, incrementally during the process of language comprehension—than our fully-formed 'offline' judgements: in their original study, Fischler et al. (1984) found that the word *bird* elicited an N400 response of smaller amplitude than *tree* in both contexts, indicating that it was more strongly predicted.

Similar effects have been reported (Kassner and Schütze, 2020; Kalouli et al., 2022) for other transformers such as Transformer-XL (Dai et al., 2019), RoBERTa (Liu et al., 2019), and ALBERT (Lan et al., 2020), as well as ELMo (Peters et al., 2018). Worse, recent work suggests that as language models increase in size, their ability to deal with negation may degrade: an inverse scaling relationship has been reported for performance at a wide range of tasks when prompts include negation (McKenzie et al., 2022b; Jang et al., 2023), though it is possible that this may reverse at extremely large scales (Wei et al., 2022).

Negation may be particularly challenging for statistical language models because its presence radically alters the meaning of a sentence, but negation occurs in only about 10% of sentences (Jiménez-Zafra et al., 2020). Quantifiers similarly impose radical modulations to meaning while also being relatively infrequent (see Appendix B). In the present study, we focus on quantifiers indicating typicality such as *most* and *few*. To the best of our knowledge, only one study has evaluated model predictions following any quantifiers (Kalouli et al., 2022), and it focused on words corresponding to logical quantifiers such as *all*, *every*, and *some*. The few studies involving the quantifiers we address either focus on predicting the quantifier itself (Pezzelle et al., 2018; Talmor et al., 2020), or use RNNs to investigate modeling significant effects on the N400 without any form of evaluation (Michaelov and Bergen, 2020). This study, therefore, represents the first attempt to explicitly evaluate the predictions of language models following *most* and *few*-type quantifiers.

In the present study, we carry out two experiments. In the first, following Ettinger (2020), we use the stimuli from a previously published N400 study (Urbach and Kutas, 2010). In it, Urbach and Kutas (2010) found that while *most* and *few*-type quantifiers do impact N400 amplitude, it is not enough to reverse predictions—*few farmers grow* ***crops*** elicits a smaller N400 response than *few farmers grow* ***worms***, indicating that *crops* was more strongly predicted than *worms*, even though experimental participants judged it to be less plausible off-line. We test whether language models show the same pattern of insensitivity towards the quantifiers that humans do in online measures. In this way, we test how closely the predictions of language models correlate with those underlying the human N400 response.

In our second experiment, we extend our study further. Experiment 1 aims to replicate the original N400 results of Urbach and Kutas (2010); however, one thing that it does not account for is that while a given complete sentence (e.g., *few farmers grow* ***crops.***) can be highly unlikely and implausible, sentences beginning with the same words may not be (for example, in the plausible sentence *few farmers grow* ***crops*** *in the winter*). Experiment 1 does not distinguish between these possibilities, and while it is important to test the sensitivity of language models to *few*-type quantifiers, if they fail to show a difference for complete sentences including the final period (e.g., *few farmers grow* ***crops.***), this is more concerning. Thus, in Experiment 2, we run the same stimuli as Experiment 1, but including a period following the final word (e.g., *crops./worms.*).

2 Experiment 1: Replication of Urbach and Kutas (2010)

2.1 Materials

In this experiment, we use all the stimuli from two experiments carried out by Urbach and Kutas (2010). These are made up of 120 sentence frames with 8 different sentence types falling into 4 experimental conditions, for a total of 960 sentences. The 4 conditions had a 2x2 design—each stimulus was either typical (T) or atypical (A), and had either

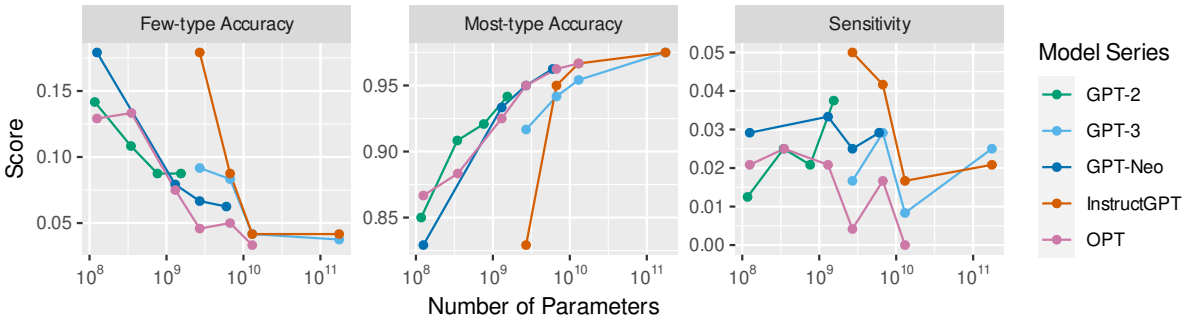

Figure 1: Accuracy and sensitivity of all models.

a *most*-type or *few*-type quantifier. An example of the 8 sentence types comprising one sentence frame is shown in (2).

(2) (a) *Most* squirrels gather **nuts**... (T, *most*)
(b) *Most* squirrels gather **nails**... (A, *most*)
(c) *Few* squirrels gather **nuts**... (T, *few*)
(d) *Few* squirrels gather **nails**... (A, *few*)
(e) Squirrels *often* gather **nuts**... (T, *most*)
(f) Squirrels *often* gather **nails**... (A, *most*)
(g) Squirrels *rarely* gather **nuts**... (T, *few*)
(h) Squirrels *rarely* gather **nails**... (A, *few*)

The quantifiers used in sentences (a)-(d) differed by sentence frame; see Appendix B for a full list.

2.2 Language Models

To cover a range of language models with different training data and numbers of parameters, we run our analyses on the GPT-2 (Radford et al., 2019), GPT-3 (Brown et al., 2020), GPT-Neo (Black et al., 2021; including GPT-J, Wang and Komatsuzaki, 2021), and OPT (Zhang et al., 2022) language models. We also include an analysis of the first series of InstructGPT models (`text-davinci-001` etc.), which were finetuned on human-written and highly-rated model-generated responses (OpenAI, 2023).

2.3 Evaluation

For each stimulus sentence, we calculate the surprisal of the critical word, that is, the word for which the N400 response was measured in the original study. Because humans only encounter the context preceding the critical word when processing the word, and because the language models we analyze are all autoregressive, we only consider the surprisal of the critical word given its preceding context. To do this we truncated the sentence before the critical word, and then used the relevant language model to calculate the probability p of the target word given the preceding context, which was then converted to surprisal S following Equation 1.

$$S = -\log p(w_i|w_1 ... w_{i-1}) \quad (1)$$

In previous work of this type (e.g., Ettinger, 2020), only words that were single tokens in the models' vocabularies were used. In this study, all models are autoregressive, so for multi-token words, consecutive sub-word tokens can be predicted, the product of which is a well-defined probability for the whole word. The surprisal of such words, then, is the sum of the surprisals of the sub-word tokens. Calculating surprisal this way allows us to compare the predictions of all the models for all the stimuli in the original experiment.

In order to evaluate how well each model takes into account the quantifier in its predictions, we compared which of the two possible critical words (typical or atypical) had a lower surprisal, i.e., was more strongly predicted by the model. To align with human plausibility judgements, following a *most*-type quantifier, the typical continuation was judged to be correct, and following a *few*-type quantifier, the atypical continuation was judged to be correct. Accuracy was calculated as the fraction of the stimulus pairs for which the model predicted the appropriate critical word—that is, predicted the correct continuation more strongly than the incorrect one. For example, the set of stimuli presented in (2) is made up of 4 pairs of stimuli, and for a model to achieve 100% accuracy (4/4), it would need to predict (a) over (b), (d) over (c), (e) over (f), and (h) over (g). This design intrinsically controls for any differences in unconditioned probability among the final words themselves.

Following Ettinger (2020), we also analyzed model sensitivity to the quantifiers. In the present study, this corresponds to the question of whether,

for a given sentence frame, the model makes a different prediction following a *few*-type quantifier than it does following a *most*-type quantifier. We defined sensitivity as the proportion of stimuli for which the model correctly predicts the critical word following both the *most*-type and the *few*-type quantifier. Thus, the stimuli in each sentence frame provide 2 data points for sensitivity: in (2), sensitivity is calculated for (a)-(d) and for (e)-(h). For the (a)-(d) stimuli, a model would be considered sensitive to the quantifier if it correctly predicted (a) over (b) *and* (d) over (c). Code and data are available at `https://osf.io/vjyw9`.

2.4 Results

Each model's accuracy at predicting the critical words following *most*- and *few*-type quantifiers is shown in Figure 1. All model series show the same general tendencies in accuracy: (1) they perform quite poorly for *few*-type quantifiers but relatively well for *most*-type quantifiers; and (2) as model size increases, word prediction following *most*-type quantifiers improves, but it degrades following *few*-type quantifiers. Figure 1 does show small exceptions to this pattern. From GPT-2 762M to 1542M and from InstructGPT 13B to 175B, while *most*-performance increases, *few*-performance does not decrease. Furthermore, from OPT 125M to 350M, and from OPT 2.7B to 6.7B, there is actually a slight improvement. Nonetheless, these differences are small compared to the overall decreases in performance, and the general trends are still clear—for example, no model performs better on *few*-type quantifiers than a model two or more sizes smaller.

With sensitivity, as shown in Figure 1, some models improve as they increase in size, and some get worse; however, even the greatest distance between the sensitivity of two models in the same series (InstructGPT 2.7B and 13B) is only 3.4%. Thus, other than the general fact that sensitivity is low across all models, there does not appear to be any clear pattern, suggesting that sensitivity does not drive the effects seen in accuracy. All accuracy and sensitivity scores can be found in Appendix A.

2.5 Discussion

These results show that contemporary autoregressive transformer models perform poorly on *few*-type quantifiers, and that as these models increase in size, they tend to improve at predicting words following *most*-type quantifiers but get worse at predicting words following *few*-type quantifiers. In fact, we see that models that better predicted the more typical word after a *most*-type quantifier were also worse at predicting the less typical word following a *least*-type quantifier. The fact that models were evaluated on which of the two options they predicted to be more likely, combined with generally poor and largely invariant sensitivity (peaking at 5%), suggests that the larger models generally made predictions increasingly in accordance with typicality, overwhelming any sensitivity to quantifier type. This aligns with previous work on negation and logical quantifiers in language models (Ettinger, 2020; Kassner and Schütze, 2020; Kalouli et al., 2022), as well as the N400 results of the original study by Urbach and Kutas (2010).

3 Experiment 2: Sentence-final nouns

3.1 Method

The models and evaluation approach were identical to Experiment 1. The materials were identical to Experiment 1 with the single difference that all nouns were followed by a period, and the surprisal of this period was included when calculating the total surprisal of the critical word (e.g., *nuts.* or *nails.* for the example presented in (2)). Thus, surprisal reflected both the surprisal of the critical word in context and the surprisal of the word being followed by a period, i.e., being the last word in the sentence. For a discussion of modeling the probability of sentence-final words in this way, see Szewczyk and Federmeier (2022).

3.2 Results

Results are shown in Figure 2. As in Experiment 1, larger models perform worse overall. However, there is a small improvement in the very largest GPT-3 and InstructGPT models relative to the second-largest models of the same type, both in *few*-type accuracy and sensitivity. Performance also increases on these metrics between OPT 2.7B and OPT 6.7B; however, this decreases with OPT 13B. All accuracy and sensitivity scores can be found in Appendix A.

3.3 Discusion

Overall, the results are similar to those of Experiment 1: Larger models of the same type perform worse than smaller models. Whether the small improvement of the largest GPT-3 and InstructGPT models relative to the second-largest models is a fluctuation like that seen for OPT or the beginnings

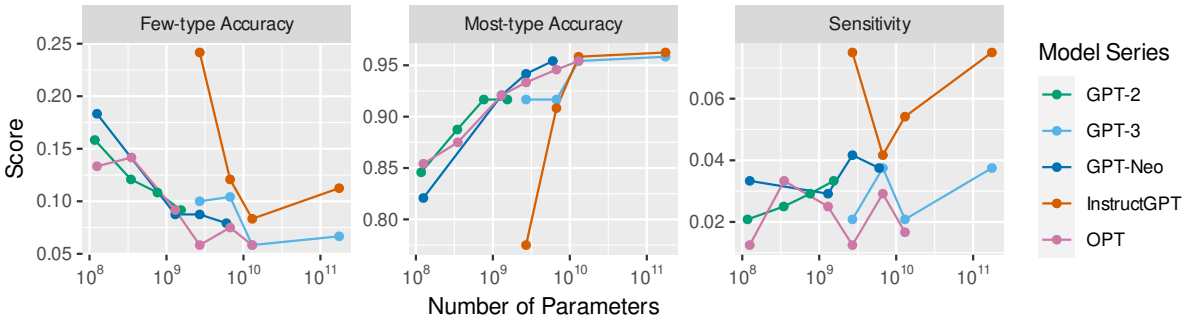

Figure 2: Accuracy and sensitivity of all models on stimuli with added periods (e.g., *Few squirrels gather* ***nuts.***).

of a U-shaped curve (see Wei et al., 2022) is a question for further research.

4 General Discussion

In this study, we investigated whether language models show the same insensitivity towards *few*-type and *most*-type quantifiers observed in the predictions made by humans during language comprehension, as indexed by the N400 response. We find that when tested on the same stimuli, they do, predicting the ostensibly implausible *few squirrels gather* ***nuts*** to be more likely than *few squirrels gather* ***nails***. Moreover, we find that as language models increase in size, they tend to show this effect to a greater extent, an example of inverse scaling. Based on our analysis of sensitivity and accuracy with *most*-type quantifiers, we hypothesize that these results are due to a low degree of sensitivity to quantifiers and an increase in sensitivity to typicality. In other words, language models appear to be increasingly sensitive to the fact that *squirrels gather* ***nuts*** is more plausible than *squirrels gather* ***nails***, but not to the effect on meaning that is caused by a preceding *most* or *few*.

It is often assumed that as models increase in size and are trained on more data, their performance on natural language tasks generally improves—indeed, evidence supports this (Brown et al., 2020; Raffel et al., 2020; Lin et al., 2021; Srivastava et al., 2022; Hoffmann et al., 2022; Rae et al., 2022; Zhang et al., 2022; Chowdhery et al., 2022). However, the predictions of larger models and those trained on more data also increasingly correlate with human incremental online predictions, in particular those indexed by N400 amplitude (Frank et al., 2015; Aurnhammer and Frank, 2019a,b; Michaelov and Bergen, 2020; Merkx and Frank, 2021; Michaelov et al., 2021, 2022). The two are often aligned—it is easier for humans to process well-formed sentences with plausible semantics (Frisch and Schlesewsky, 2005; Nieuwland et al., 2020). But in cases such as the present study, the two are not aligned, and we see instead that the predictions of larger models correlate better with human online predictions, even when these are contrary to offline judgements. Thus, the increased performance we see at tasks corresponding to offline human judgements—and note that virtually all manually-annotated tasks are based on offline human judgements—may in fact be a by-product of the models' predictions resembling the online predictions.

Fortunately, the literature boasts a wealth of psycholinguistic studies where metrics of online prediction such as the N400 appear to conflict with offline judgements. Future work could use these to identify phenomena where language models may struggle to make predictions in line with human judgements. Such cases are important to detect as use of LMs becomes more widespread. But by the same token, the present study shows that as language models increase in size, even when augmented by finetuning on desirable responses, they can make predictions that align less and less with explicit human judgements.

This may be a clear indication of an inherent 'outer misalignment' present in language models: while humans might like language models to generate plausible sentences, by their nature they can only generate the most statistically probable ones. Just as there is no guarantee of accuracy or coherence (Bender et al., 2021), there is no guarantee of plausibility. While it may be possible to tailor training to avoid specific known issues, this misalignment between probability and plausibility may pose a fundamental challenge with current approaches that aim to use language models as general-purpose natural language systems.

Limitations

There are two main limitations to our study. The first is that the stimuli used were limited to those provided by Urbach and Kutas's (2010) study. This is because, as stated, we wanted to be able to compare the patterns in the language models' predictions to the patterns in the human N400 response. Thus, we do not look at logical quantifiers like Kalouli et al. (2022), or any others that have previously been studied (in, e.g., Pezzelle et al., 2018; Talmor et al., 2020).

The other (and perhaps more important) limitation is in the models we were able to use. Crucially, we were not able to access models larger than GPT-3 175B such as PaLM 540B (Chowdhery et al., 2022). This is important because recent work has shown that some inverse scaling patterns become U-shaped (i.e., as language model size increases, performance degrades and then improves again) with such larger models (Wei et al., 2022).

Ethics Statement

Our work complies with the ACL Ethics Policy. Beyond this, we are not aware of any way in which the results of this study may be harmful—in fact, if anything, identifying the limitations of large language models is something that is likely to reduce possible harms by demonstrating cases where their use is not suitable.

From an environmental perspective, we did not train any models; we only used pretrained models for analysis, limiting energy consumption. With the exception of the GPT-3 and InstructGPT models and OPT 13B, all analyses were run on an NVIDIA RTX A6000 GPU, taking a total of 43 minutes. OPT 13B was too large to run on this GPU, and thus was run on an Intel Dual Xeon E7-4870 CPU for a total of 22 hours and 39 minutes. Finally, the GPT-3 and the InstructGPT models were run using the OpenAI API, and thus we do not have access to information about the GPUs used.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments. We would also like to acknowledge the other members of the Language and Cognition Lab at UCSD for their valuable discussion, as well as Roger Levy and attendees of the MIT Computational Psycholinguistics Laboratory meeting. Finally, we would like to thank the San Diego Social Sciences Computing Facility Team for the use of the Social Sciences Research and Development Environment (SSRDE) cluster. The RTX A6000 used for this research was donated by the NVIDIA Corporation.

References

Christoph Aurnhammer and Stefan L. Frank. 2019a. Comparing Gated and Simple Recurrent Neural Network Architectures as Models of Human Sentence Processing. In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society (CogSci 2019)*.

Christoph Aurnhammer and Stefan L. Frank. 2019b. Evaluating information-theoretic measures of word prediction in naturalistic sentence reading. *Neuropsychologia*, 134:107198.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, New York, NY, USA. Association for Computing Machinery.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large scale autoregressive language modeling with mesh-tensorflow. Zenodo.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr,

Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the opportunities and risks of foundation models. *ArXiv*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

Allyson Ettinger. 2020. What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Ira Fischler, Paul A. Bloom, Donald G. Childers, A. Antonio Arroyo, and Nathan W. Perry. 1984. Brain potentials during sentence verification: Late negativity and long-term memory strength. *Neuropsychologia*, 22(5):559–568.

Stefan L. Frank, Leun J. Otten, Giulia Galli, and Gabriella Vigliocco. 2015. The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, 140:1–11.

Stefan Frisch and Matthias Schlesewsky. 2005. The resolution of case conflicts from a neurophysiological perspective. *Cognitive Brain Research*, 25(2):484–498.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training Compute-Optimal Large Language Models.

Joel Jang, Seonghyeon Ye, and Minjoon Seo. 2023. Can Large Language Models Truly Understand Prompts? A Case Study with Negated Prompts. In *Proceedings of The 1st Transfer Learning for Natural Language Processing Workshop*, pages 52–62. PMLR.

Salud María Jiménez-Zafra, Roser Morante, María Teresa Martín-Valdivia, and L. Alfonso Ureña-López. 2020. Corpora Annotated with Negation: An Overview. *Computational Linguistics*, 46(1):1–52.

Aikaterini-Lida Kalouli, Rita Sevastjanova, Christin Beck, and Maribel Romero. 2022. Negation, Coordination, and Quantifiers in Contextualized Language Models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3074–3085, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Nora Kassner and Hinrich Schütze. 2020. Negated and Misprimed Probes for Pretrained Language Models: Birds Can Talk, But Cannot Fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference on Learning Representations*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2021. Few-shot Learning with Multilingual Language Models. *arXiv:2112.10668 [cs]*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019.

RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*.

Ian McKenzie, Alexander Lyzhov, Alicia Parrish, Ameya Prabhu, Aaron Mueller, Najoung Kim, Sam Bowman, and Ethan Perez. 2022a. The inverse scaling prize.

Ian McKenzie, Alexander Lyzhov, Alicia Parrish, Ameya Prabhu, Aaron Mueller, Najoung Kim, Sam Bowman, and Ethan Perez. 2022b. Inverse scaling prize: First round winners.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer Sentinel Mixture Models. In *International Conference on Learning Representations*.

Danny Merkx and Stefan L. Frank. 2021. Human Sentence Processing: Recurrence or Attention? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 12–22, Online. Association for Computational Linguistics.

James A. Michaelov, Megan D. Bardolph, Seana Coulson, and Benjamin K. Bergen. 2021. Different kinds of cognitive plausibility: Why are transformers better than RNNs at predicting N400 amplitude? In *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*, pages 300–306, University of Vienna, Vienna, Austria (Hybrid).

James A. Michaelov and Benjamin K. Bergen. 2020. How well does surprisal explain N400 amplitude under different experimental conditions? In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 652–663, Online. Association for Computational Linguistics.

James A. Michaelov, Seana Coulson, and Benjamin K. Bergen. 2022. So Cloze yet so Far: N400 Amplitude is Better Predicted by Distributional Information than Human Predictability Judgements. *IEEE Transactions on Cognitive and Developmental Systems*.

Mante S. Nieuwland, Dale J. Barr, Federica Bartolozzi, Simon Busch-Moreno, Emily Darley, David I. Donaldson, Heather J. Ferguson, Xiao Fu, Evelien Heyselaar, Falk Huettig, E. Matthew Husband, Aine Ito, Nina Kazanina, Vita Kogan, Zdenko Kohút, Eugenia Kulakova, Diane Mézière, Stephen Politzer-Ahles, Guillaume Rousselet, Shirley-Ann Rueschemeyer, Katrien Segaert, Jyrki Tuomainen, and Sarah Von Grebmer Zu Wolfsthurn. 2020. Dissociable effects of prediction and integration during language comprehension: Evidence from a large-scale study using brain potentials. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 375(1791):20180522.

OpenAI. 2023. Model index for researchers.

Ethan Perez, Ian McKenzie, and Sam Bowman. 2022. Announcing the Inverse Scaling Prize ($250k Prize Pool).

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Sandro Pezzelle, Shane Steinert-Threlkeld, Raffaella Bernardi, and Jakub Szymanik. 2018. Some of Them Can be Guessed! Exploring the Effect of Linguistic Context in Predicting Quantifiers. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 114–119, Melbourne, Australia. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. page 24.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2022. Scaling Language Models: Methods, Analysis & Insights from Training Gopher.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021. AI and the Everything in the Whole Wide World Benchmark.

Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, 1.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramón Risco Delgado, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Bider-

man, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Timothy Telleen-Lawton, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2022. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models.

Jakub M. Szewczyk and Kara D. Federmeier. 2022. Context-based facilitation of semantic access follows both logarithmic and linear functions of stimulus probability. *Journal of Memory and Language*, 123:104311.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. oLMpics-On What Language Model Pre-training Captures. *Transactions of the Association for Computational Linguistics*, 8:743–758.

Thomas P. Urbach and Marta Kutas. 2010. Quantifiers more or less quantify on-line: ERP evidence for partial incremental interpretation. *Journal of Memory and Language*, 63(2):158–179.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 billion parameter autoregressive language model.

Jason Wei, Yi Tay, and Quoc V. Le. 2022. Inverse scaling can become U-shaped.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: Open Pre-trained Transformer Language Models.

A Scores

The performance of each model is presented in Table 1.

	Critical word			Critical word + period		
	Accuracy		Sens.	Accuracy		Sens.
Model	***most***	***few***		***most***	***few***	
GPT-2 117M (`gpt2`)	0.850	0.142	0.013	0.846	0.158	0.021
GPT-2 345M (`gpt2-medium`)	0.908	0.108	0.025	0.887	0.121	0.025
GPT-2 762M (`gpt2-large`)	0.921	0.088	0.021	0.917	0.108	0.029
GPT-2 1542M (`gpt2-xl`)	0.942	0.088	0.038	0.917	0.092	0.033
GPT-3 2.7B (`ada`)	0.917	0.092	0.017	0.917	0.1	0.021
GPT-3 6.7B (`babbage`)	0.942	0.083	0.029	0.917	0.104	0.038
GPT-3 13B (`curie`)	0.954	0.042	0.008	0.954	0.058	0.021
GPT-3 175B (`davinci`)	0.975	0.038	0.025	0.958	0.067	0.038
InstructGPT 2.7B (`text-ada-001`)	0.829	0.179	0.050	0.775	0.242	0.075
InstructGPT 6.7B (`text-babbage-001`)	0.950	0.088	0.042	0.908	0.121	0.042
InstructGPT 13B (`text-curie-001`)	0.967	0.042	0.017	0.958	0.083	0.054
InstructGPT 175B (`text-davinci-001`)	0.975	0.042	0.021	0.963	0.112	0.075
GPT-Neo 125M (`EleutherAI/gpt-neo-125m`)	0.829	0.179	0.029	0.821	0.183	0.033
GPT-Neo 1.3B (`EleutherAI/gpt-neo-1.3B`)	0.933	0.079	0.033	0.921	0.088	0.029
GPT-Neo 2.7B (`EleutherAI/gpt-neo-2.7B`)	0.950	0.067	0.025	0.942	0.088	0.042
GPT-J 6B (`EleutherAI/gpt-j-6b`)	0.963	0.062	0.029	0.954	0.079	0.038
OPT 125M (`facebook/opt-125m`)	0.867	0.129	0.021	0.854	0.133	0.013
OPT 350M (`facebook/opt-350m`)	0.883	0.133	0.025	0.875	0.142	0.033
OPT 1.3B (`facebook/opt-1.3b`)	0.925	0.075	0.021	0.921	0.092	0.025
OPT 2.7B (`facebook/opt-2.7b`)	0.950	0.046	0.004	0.933	0.058	0.013
OPT 6.7B (`facebook/opt-6.7b`)	0.963	0.050	0.017	0.946	0.075	0.029
OPT 13B (`facebook/opt-13b`)	0.967	0.033	0	0.954	0.058	0.017

Table 1: Accuracy and sensitivity scores for all models.

B Quantifiers

Table 2 lists all quantifiers used and the proportion of sentences in WikiText-103 that contain them.

	***Most*-type**		***Few*-type**	
	Quantifier	**Frequency (sentences)**	**Quantifier**	**Frequency (sentences)**
	most	0.025177	few	0.005870
	almost all	0.000305	almost no	0.000098
	practically all	0.000009	practically no	0.000008
	a large number of	0.000300	a small number of	0.000131
	nearly all	0.000170	rather few	0.000001
	lots of	0.000153	hardly any	0.000017
	a lot of	0.000745	a very few	0.000010
	many	0.015874	few	0.005870
	often	0.005766	rarely	0.000610
Total		0.046809		0.006717

Table 2: In each sentence frame, *most* and *few*-type quantifiers were matched based on their meanings as length in number of words (Urbach and Kutas, 2010). Matched quantifiers are shown beside each other. As can be seen, *few* is matched to both *most* and *many*. The frequency of each quantifier is given in terms of the proportion of sentences in WikiText-103 (Merity et al., 2017) that contain it. The total frequencies are the number of sentences in WikiText-103 that contain at least one of either the *few*-type or *most*-type quantifiers; not the sum of the individual quantifier frequencies.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations

☑ A2. Did you discuss any potential risks of your work?
Ethics Statement

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract, 1 (Introduction)

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

1 (Introduction), 2 (Method), 3 (Results)

☑ B1. Did you cite the creators of artifacts you used?
1 (Introduction), 2 (Experiment 1)

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Supplementary Materials (OSF repository linked in paper)

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Supplementary Materials (OSF repository linked in paper)

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. We only used the stimuli constructed by researchers from a previously-published study (Urbach and Kutas, 2010)

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Abstract, 1 (Introduction), 2 (Experiment 1), 3 (Experiment 2)

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Discussed in section 2.1 (Materials)

C ☑ Did you run computational experiments?

2 (Experiment 1), 3 (Experiment 2)

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Parameters: Appendix; Computational budget and infrastructure: Ethics Statement

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
2 (Experiment 1), 3 (Experiment 2)

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
2 (Experiment 1), 3 (Experiment 2)

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
All data, code for running the models, and analyses are provided as supplementary materials in an OSF repository (link in paper).

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.