OPT: Open Pre-trained Transformer Language Models

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**, **Luke Zettlemoyer**

Meta AI

{susanz,roller,naman}@fb.com

Abstract

Large language models, which are often trained for hundreds of thousands of compute days, have shown remarkable capabilities for zero- and few-shot learning. Given their computational cost, these models are difficult to replicate without significant capital. For the few that are available through APIs, no access is granted to the full model weights, making them difficult to study. We present Open Pre-trained Transformers (OPT), a suite of decoder-only pre-trained transformers ranging from 125M to 175B parameters, which we aim to fully and responsibly share with interested researchers. We show that OPT-175B is comparable to GPT-3,[1] while requiring only 1/7th the carbon footprint to develop. We are also releasing our logbook detailing the infrastructure challenges we faced, along with code for experimenting with all of the released models.

1 Introduction

Large language models (LLMs) trained on massive text collections have shown surprising emergent capabilities to generate text and perform zero- and few-shot learning (Brown et al., 2020; Lieber et al., 2021; Smith et al., 2022; Rae et al., 2021; Chowdhery et al., 2022). While in some cases the public can interact with these models through paid APIs, full model access is currently limited to only a few highly resourced labs.[2] This restricted access has limited researchers' ability to study how and why these large language models work, hindering progress on improving known challenges in areas such as robustness, bias, and toxicity.

In this technical report, we present Open Pre-trained Transformers (OPT), a suite of decoder-only pre-trained transformers ranging from 125M to 175B parameters, which we aim to fully and responsibly share with interested researchers. We train the OPT models to roughly match the performance and sizes of the GPT-3 class of models, while also applying the latest best practices in data collection and efficient training. Our aim in developing this suite of OPT models is to enable reproducible and responsible research at scale, and to bring more voices to the table in studying the impact of these LLMs. Definitions of risk, harm, bias, and toxicity, etc., should be articulated by the collective research community as a whole, which is only possible when models are available for study.

We are releasing all of our models between 125M and 66B parameters, and will provide full research access to OPT-175B upon request. Access will be granted to academic researchers; those affiliated with organizations in government, civil society, and academia; and those in industry research laboratories. We are also releasing both the logbook of our model creation as well as our codebase, metaseq,[3] which enabled training OPT-175B on 992 80GB A100 GPUs, reaching 147 TFLOP/s utilization per GPU. From this implementation, and from using the latest generation of NVIDIA hardware, we are able to develop OPT-175B using only 1/7th the carbon footprint of GPT-3. While this is a significant achievement, the energy cost of creating such a model is still nontrivial, and repeated efforts to replicate a model of this size will only amplify the growing compute footprint of these LLMs.

We believe the entire AI community — academic researchers, civil society, policymakers, and industry — must work together to develop clear

[*] Equal contribution.

[†] Work done while at Meta AI.

[1] Following Brown et al. (2020), we use GPT-3 to refer to both the 175B model and the smaller scale models as well.

[2] Exceptions include work by EleutherAI, who released dense models up to 20B in size (Black et al., 2022), Salesforce (Nijkamp et al., 2022), and Meta AI, who released dense models up to 13B and sparse models up to 1.1T (Artetxe et al., 2021). There is also ongoing work from the BigScience workshop (https://bigscience.huggingface.co/), which aims to open source very large multilingual language models and datasets.

[3] https://github.com/facebookresearch/metaseq

Model	#L	#H	d_{model}	LR	Batch
125M	12	12	768	$6.0e{-4}$	0.5M
350M	24	16	1024	$3.0e{-4}$	0.5M
1.3B	24	32	2048	$2.0e{-4}$	1M
2.7B	32	32	2560	$1.6e{-4}$	1M
6.7B	32	32	4096	$1.2e{-4}$	2M
13B	40	40	5120	$1.0e{-4}$	4M
30B	48	56	7168	$1.0e{-4}$	4M
66B	64	72	9216	$0.8e{-4}$	2M
175B	96	96	12288	$1.2e{-4}$	2M

Table 1: **Model architecture details.** We report the number of layers (#L), number of attention heads (#H), and the embedding size (d_{model}). We also report the peak Learning Rate (LR) and global batch size in number of tokens (Batch).

guidelines around responsible AI in general and responsible LLMs in particular, given their centrality in many downstream language applications. A much broader segment of the AI community needs access to these models in order to conduct reproducible research and collectively drive the field forward. With the release of OPT-175B and smaller-scale baselines, we hope to increase the diversity of voices defining the ethical considerations of such technologies.

2 Method

2.1 Models

We present results on eight Transformer language models ranging from 125 million to 175 billion parameters. Architectural details are displayed in Table 1. In the interest of transparency, and to reduce risk of training instabilities, our models and hyperparameters largely follow Brown et al. (2020), with variations in batch size mostly to obtain increased computational efficiency.

2.2 Training Setup

For weight initialization, we follow the same settings provided in the Megatron-LM codebase,[4] using a normal distribution with zero mean and standard deviation of 0.006. Standard deviation for output layers are scaled by a $1.0/\sqrt{2L}$ term where L is the total number of layers. All bias terms are initialized as 0, and all models are trained with ReLU activation and a sequence length of 2048.

[4] https://github.com/NVIDIA/Megatron-LM/blob/main/examples/pretrain_gpt3_175B.sh

We use an AdamW optimizer (Loshchilov and Hutter, 2017) with (β_1, β_2) set to $(0.9, 0.95)$, and weight decay of 0.1. We follow a linear learning rate schedule, warming up from 0 to the maximum learning rate over the first 2000 steps in OPT-175B, or over 375M tokens in our smaller baselines, and decaying down to 10% of the maximum LR over 300B tokens. A number of mid-flight changes to LR were also required (see Section 2.5). Our batch sizes range from 0.5M to 4M depending on the model size (see Table 1) and is kept constant throughout the course of training.

We use a dropout of 0.1 throughout, but we do not apply any dropout to embeddings. We clip gradient norms at 1.0, except for some mid-flight changes that reduce this threshold down from 1.0 to 0.3 (see Section 2.5). We also include a gradient predivide factor to reduce the risk of over/underflows when computing the gradient across all ranks (splitting the division by the world size of N into two division operations by $\sqrt{N}$).

2.3 Pre-training Corpus

The pre-training corpus contains a concatenation of datasets used in RoBERTa (Liu et al., 2019b), the Pile (Gao et al., 2021a), and PushShift.io Reddit (Baumgartner et al., 2020; Roller et al., 2021). All corpora were previously collected or filtered to contain predominantly English text, but a small amount of non-English data is still present within the corpus via CommonCrawl.

We removed duplicated documents across all datasets by filtering out documents via MinhashLSH (Rajaraman and Ullman, 2011) with a Jaccard similarity $\geq .95$. We found the Pile was particularly full of duplicate documents, and advise future researchers using the Pile to perform additional de-duplication processing.

We tokenize all corpora using the GPT-2 byte level BPE tokenizer (Sennrich et al., 2016; Radford et al., 2019; Brown et al., 2020). Our final corpus contains roughly 180B tokens.

RoBERTa We included the BookCorpus (Zhu et al., 2015) and Stories (Trinh and Le, 2018) subsets of the RoBERTa corpus and utilized an updated version of CCNews, containing news stories crawled through September 28, 2021. This CCNews v2 corpus was preprocessed the same way as the original RoBERTa CCNews (Liu et al., 2019b).

The Pile We included a subset of the Pile (Gao et al., 2021a), including: CommonCrawl,

DM Mathematics, Project Gutenberg, HackerNews, OpenSubtitles, OpenWebText2, USPTO and Wikipedia. Other subsets of the Pile were eliminated as we found they increased the risk of instabilities, as measured by tendency to cause spikes in gradient norms at the 1.3B scale, or were otherwise deemed unsuitable. All subsets went through additional ad-hoc whitespace normalization.

PushShift.io Reddit We included a subset of the Pushshift.io corpus produced by Baumgartner et al. (2020) and previously used by Roller et al. (2021). To convert the conversational trees into language-model-accessible documents, we extracted the longest chain of comments in each thread and discarded all other paths in the tree. This reduced the corpus by about 66%.

2.4 Training Efficiency

We trained OPT-175B on 992 80GB A100 GPUs, by utilizing Fully Sharded Data Parallel (Artetxe et al., 2021) with Megatron-LM Tensor Parallelism (Shoeybi et al., 2019). We achieve utilization of up to 147 TFLOP/s per GPU. We keep Adam state in FP32, since we shard it across all hosts, while the model weights remained in FP16. To avoid underflows, we used dynamic loss scaling, as described in Micikevicius et al. (2017).

2.5 Training Processes

Here we describe significant training process adjustments that arose during OPT-175B pre-training.

Hardware Failures We faced a significant number of hardware failures in our compute cluster while training OPT-175B. In total, hardware failures contributed to at least 35 manual restarts and the cycling of over 100 hosts over the course of 2 months. During manual restarts, the training run was paused, and a series of diagnostics tests were conducted to detect problematic nodes. Flagged nodes were then cordoned off and training was resumed from the last saved checkpoint. Given the difference between the number of hosts cycled out and the number of manual restarts, we estimate 70+ automatic restarts due to hardware failures.

Loss Divergences Loss divergences were also an issue in our training run. When the loss diverged, we found that lowering the learning rate and restarting from an earlier checkpoint allowed for the job to recover and continue training. We noticed a correlation between loss divergence, our dynamic loss scalar crashing to 0, and the l^2-norm of the activations of the final layer spiking. These observations led us to pick restart points for which our dynamic loss scalar was still in a "healthy" state ($\geq$ 1.0), and after which our activation norms would trend downward instead of growing unboundedly. Our empirical LR schedule is shown in Figure 1. Early in training, we also noticed that lowering gradient clipping from 1.0 to 0.3 helped with stability; see our released logbook for exact details. Figure 2 shows our validation loss with respect to training iterations.

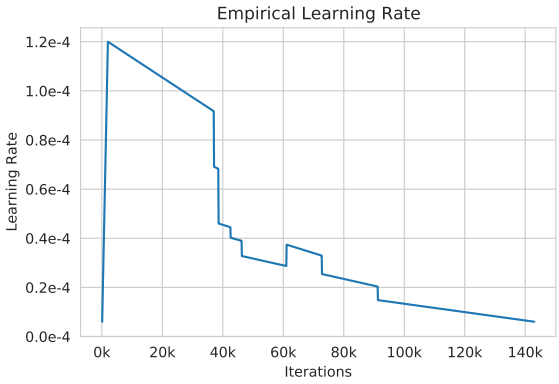

Figure 1: **Empirical LR schedule.** We found that lowering learning rate was helpful for avoiding instabilities.

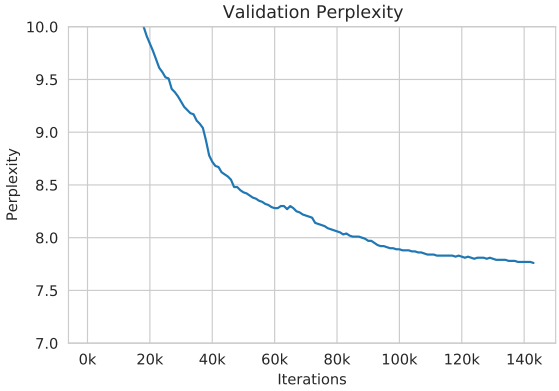

Figure 2: **Validation Perplexity.** Our mid-flight LR changes had clear effects on validation perplexity.

Other Mid-flight Changes We conducted a number of other experimental mid-flight changes to handle loss divergences. These included: switching to vanilla SGD (optimization plateaued quickly, and we reverted back to AdamW); resetting the dynamic loss scalar (this helped recover some but not all divergences); and switching to a newer version of Megatron (this reduced pressure on activation norms and improved throughput).

3 Evaluations

3.1 Prompting & Few-Shot

We evaluate our model on 16 standard NLP tasks utilized in the literature: HellaSwag (Zellers et al., 2019), StoryCloze (Mostafazadeh et al., 2016), PIQA (Bisk et al., 2020), ARC Easy and Challenge (Clark et al., 2018), OpenBookQA (Mihaylov et al., 2018), WinoGrad (Levesque et al., 2011), WinoGrande (Sakaguchi et al., 2020), and SuperGLUE (Wang et al., 2019). We follow GPT-3 (Brown et al., 2020) by using their prompts and overall experimental setup. We compare primarily to GPT-3, having aimed to re-implement their evaluation settings, but include reported performance of other LLMs on a per-task basis when available (Lieber et al., 2021; Rae et al., 2021; Hoffmann et al., 2022; Black et al., 2022)

We report performance in accuracy (omitting F1 for MultiRC and ReCoRD for consistency in evaluation metrics). For the Winograd Schema Challenge (WSC) task in the SuperGLUE benchmark, we follow (Brown et al., 2020) and formulate the task as multiple choice questions, which is known to affect performance (Liu et al., 2020).

Zero-shot Overall average zero-shot performance across all 14 tasks may be seen in Figure 3. Overall, we see our average performance follows the trend of GPT-3. However, performance can vary radically across the tasks: for a full breakdown, see Appendix A. Note that we intentionally removed MultiRC and WIC from these averages, as these datasets seem to systematically favor GPT-3 or OPT disproportionately.

Our performance roughly matched GPT-3 for 10 tasks, and underperformed in 3 tasks (ARC Challenge and MultiRC). In 3 tasks (CB, BoolQ, WSC), we find both GPT and OPT models display unpredictable behavior with respect to scale, likely due to the small size of the validation set in these 3 tasks (56, 277, and 104 examples, respectively). In WIC, we see that the OPT models always outperform the GPT-3 models, though the numbers reported by Brown et al. (2020) also seem questionable, given WIC being a binary classification task.[5] For MultiRC, we are unable to replicate the GPT-3 results using the Davinci API[6] within our evaluation setup, suggesting differences in the methods

[5] Brown et al. (2020) reports 0% accuracy on WIC, which implies 100% accuracy if the classification was inverted.

[6] https://beta.openai.com/docs/engines/overview

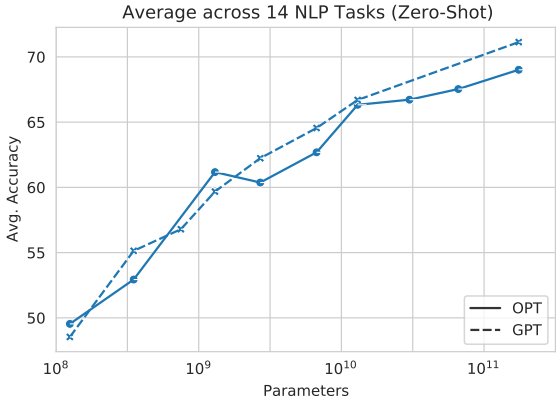

Figure 3: **Zero-shot NLP Evaluation Averages**. Across a variety of tasks and model sizes, OPT largely matches the reported averages of GPT-3. However, performance varies greatly per task: see Appendix A.

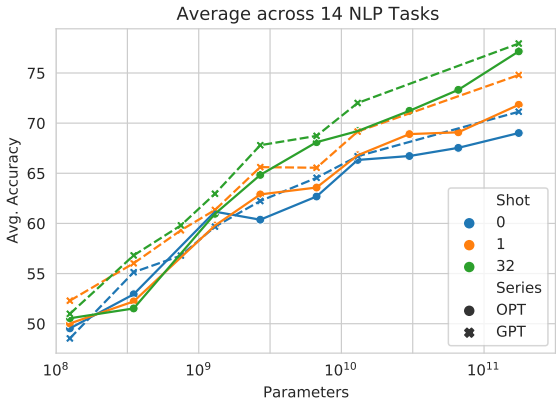

Figure 4: **Multi-shot performance**. OPT performance for one- and few-shot lags behind GPT-3 models, but performance depends heavily per task; see Appendix A.

of evaluation on this task. For BoolQ and WSC, we note that both OPT and GPT models seem to hover around majority-class accuracy, suggesting small perturbations in probability masses may be dominating the evaluations.

Chinchilla (Hoffmann et al., 2022) and Gopher (Rae et al., 2021) perform roughly consistently with others for their parameter sizes, while PaLM (Chowdhery et al., 2022) generally performs better across all settings, even when controlling for number of parameters. We speculate the high performance of PaLM comes predominantly from higher quality and diversity of pre-training data.

One-shot and Few-shot Average multi-shot in-context performance is shown in Figure 4 (again, omitting MultiRC and WIC), with detailed performances shown in Appendix A. Across the average

of all metrics, we find that OPT models perform similarly to GPT-3 models. However, as with zero-shot, breaking down these results per task shows a different story: in the same set of 10 datasets as zero-shot, we see similar performance across the two models. Some of the remaining datasets show inconsistent performance with respect to model size for both OPT and GPT-3 models (BoolQ, CB, WSC, RTE). In MultiRC, we consistently see underperformance of OPT models compared to GPT-3 models. Similar to our zero-shot evaluation, we hypothesize our one- and few-shot evaluation setup may differ significantly from Brown et al. (2020).

3.2 Dialogue

Given that LLMs are known to be an integral component of modern dialogue models (Adiwardana et al., 2020; Roller et al., 2021; Thoppilan et al., 2022; Rae et al., 2021; Chowdhery et al., 2022), we additionally evaluate OPT-175B on several open source dialogue datasets. In particular, we follow Roller et al. (2021), and evaluate on ConvAI2 (Dinan et al., 2020b), Wizard of Wikipedia (Dinan et al., 2019b), Empathetic Dialogues (Rashkin et al., 2019), and Blended Skill Talk (Smith et al., 2020). We additionally evaluate on the more recent Wizard of Internet dataset (Komeili et al., 2021). We focus our comparisons primarily against existing open source dialogue models including the fine-tuned BlenderBot 1 (Roller et al., 2021) and its pre-training counterpart Reddit 2.7B. We also compare against the fine-tuned R2C2 BlenderBot, a 2.7B parameter BlenderBot-like model trained by Shuster et al. (2022).

We report Perplexity and Unigram F1 (UF1) overlap, following the metrics of the ConvAI2 competition (Dinan et al., 2020b). To control for different tokenization in each of the models, we normalize all perplexities to be in the space of the GPT-2 tokenizer (Radford et al., 2019). We also note which models are supervised with respect to these dialogue tasks and which are unsupervised. For OPT-175B, all generations are performed using greedy decoding up to a maximum of 32 tokens. We do not attempt to prompt the model at all except for alternating "Person 1:" and "Person 2:" lines of dialogue. The remaining models use the generation parameters found in BlenderBot 1.

Results are shown in Table 2. We see that OPT-175B significantly outperforms the also-unsupervised Reddit 2.7B model on all tasks, and performs competitively with the fully supervised BlenderBot 1 model, especially in the ConvAI2 dataset. On the Wizard-of-Internet dataset, which is fully unsupervised for all models, we see that OPT-175B obtains the lowest perplexity but still has lower UF1 than the models with Wizard-of-Wikipedia supervision.

We were somewhat surprised that the evaluations of the unsupervised OPT-175B model were as competitive as BlenderBot 1 on the ConvAI2 dataset. This may indicate leakage of the ConvAI2 dataset into the general pre-training corpus or even into the validation data as evaluated in Table 2. To address concerns of leakage, we searched our pre-training corpus for the first conversation in the ConvAI2 dataset, but we did not find any overlap. We additionally evaluated OPT-175B on the ConvAI2 hidden test set, which has never been publicly released, and achieved 10.7 ppl and .185 UF1, matching the performance of the validation set. Furthermore, we evaluated OPT-175B on a subset of the ConvAI2-like MultiSessionChat (MSC) dataset (Xu et al., 2021b) and obtained a perplexity of 9.7 and UF1 of .177, indicating the model is generalizing well across multiple PersonaChat-like datasets. Since both MSC and WoI datasets were released *after* the CommonCrawl snapshot used in pre-training corpus, there is minimal risk of leakage. We conclude that OPT-175B has a strong ability to maintain a consistent persona across conversations, a behavior also highlighted in LaMDA (Thoppilan et al., 2022).

4 Bias & Toxicity Evaluations

To understand the potential harm of OPT-175B, we evaluate a series of benchmarks related to hate speech detection, stereotype awareness, and toxic content generation. While there may be shortcomings in these benchmarks (Blodgett et al., 2021; Jacobs and Wallach, 2021), these measurements provide a first step towards understanding the limitations of OPT-175B. We compare primarily against GPT-3 Davinci, as these benchmarks were not yet available to be included in Brown et al. (2020).

4.1 Hate Speech Detection

Using the ETHOS dataset provided in Mollas et al. (2020) and instrumented by Chiu and Alexander (2021), we measure the ability of OPT-175B to identify whether or not certain English statements are racist or sexist (or neither). In the zero-, one-,

Model	Eval	Perplexity (↓)					Unigram F1 (↑)				
		C2	WW	ED	BST	WoI	C2	WW	ED	BST	WoI
Reddit 2.7B	Unsup.	18.9	21.0	11.6	17.4	18.0	.126	.133	.135	.133	.124
BlenderBot 1	Sup.	**10.2**	12.5	**9.0**	11.9	14.7	.183	.189	.192	.178	.154
R2C2 BlenderBot	Sup.	10.5	**12.4**	9.1	**11.7**	14.6	**.205**	**.198**	**.197**	**.186**	**.160**
OPT-175B	Unsup.	10.8	13.3	10.3	12.1	**12.0**	.185	.152	.149	.162	.147

Table 2: **Dialogue Evaluations.** OPT-175B, in a fully unsupervised setting, performs competitively against fully supervised models.

Setup	Davinci	OPT-175B
Zero-shot	.628	**.667**
One-shot	.616	**.713**
Few-shot (binary)	.354	**.759**
Few-shot (multiclass)	.672	**.812**

Table 3: **Hate speech detection.** F1 scores of detecting hate speech between Davinci and OPT-175B. OPT-175B considerably outperforms Davinci in all settings.

and few-shot binary cases, the model is presented with text and asked to consider whether the text is racist or sexist and provide a yes/no response. In the few-shot multiclass setting, the model is asked to provide a yes/no/neither response.

Results are presented in Table 3. With all of our one-shot through few-shot configurations, OPT-175B performs considerably better than Davinci. We speculate this occurs from two sources: (1) evaluating via the Davinci API may be bringing in safety control mechanisms beyond the original 175B GPT-3 model used in Brown et al. (2020); and (2) the significant presence of unmoderated social media discussions in the pre-training dataset has provided additional inductive bias to aid in such classification tasks.

4.2 CrowS-Pairs

Developed for masked language models, CrowS-Pairs (Nangia et al., 2020) is a crowdsourced benchmark aiming to measure intrasentence level biases in 9 categories: gender, religion, race/color, sexual orientation, age, nationality, disability, physical appearance, and socioeconomic status. Each example consists of a pair of sentences representing a stereotype, or anti-stereotype, regarding a certain group, with the goal of measuring model preference towards stereotypical expressions. Higher scores indicate higher bias exhibited by a model.

Category	GPT-3	OPT-175B
Gender	**62.6**	65.7
Religion	73.3	**68.6**
Race/Color	**64.7**	68.6
Sexual orientation	**76.2**	78.6
Age	**64.4**	67.8
Nationality	**61.6**	62.9
Disability	**76.7**	**76.7**
Physical appearance	**74.6**	76.2
Socioeconomic status	**73.8**	76.2
Overall	**67.2**	69.5

Table 4: **CrowS-Pairs evaluation.** Lower is better for all categories, indicating more fairness. The OPT-175B model performs worse than Davinci in most categories.

When compared with Davinci in Table 4, OPT-175B appears to exhibit more stereotypical biases in almost all categories except for religion. Again, this is likely due to differences in training data; Nangia et al. (2020) showed that Pushshift.io Reddit corpus has a higher incidence rate for stereotypes and discriminatory text than other corpora (e.g. Wikipedia). Given this is a primary data source for OPT-175B, the model may have learned more discriminatory associations, which directly impacts its performance on CrowS-Pairs.

4.3 StereoSet

Following Lieber et al. (2021) and Artetxe et al. (2021), we use StereoSet (Nadeem et al., 2021) to measure stereotypical bias across 4 categories: profession, gender, religion, and race. In addition to intrasentence measurement (similar to CrowS-Pairs), StereoSet includes measurement at the intersentence level to test a model's ability to incorporate additional context. To account for a potential trade-off between bias detection and language modeling capability, StereoSet includes two metrics:

Category		Davinci	OPT-175B
Prof.	LMS (↑)	**78.4**	74.1
	SS (↓)	63.4	**62.6**
	ICAT (↑)	**57.5**	55.4
Gend.	LMS (↑)	**75.6**	74.0
	SS (↓)	66.5	**63.6**
	ICAT (↑)	50.6	**53.8**
Reli.	LMS (↑)	80.8	**84.0**
	SS (↓)	**59.0**	**59.0**
	ICAT (↑)	66.3	**68.9**
Race	LMS (↑)	**77.0**	74.9
	SS (↓)	57.4	**56.8**
	ICAT (↑)	**65.7**	64.8
Overall	LMS (↑)	**77.6**	74.8
	SS (↓)	60.8	**59.9**
	ICAT (↑)	**60.8**	60.0

Table 5: **StereoSet Evaluations**. Davinci and OPT-175B perform similarly across all evaluations.

Language Modeling Score (LMS) and Stereotype Score (SS), which are then combined to form the Idealized Context Association Test score (ICAT). Unlike Lieber et al. (2021), we normalize scores by token count, rather than character count, which they report improves metrics for several models.

Results are shown in Table 5. We see that Davinci and OPT-175B exhibit similar scores on aggregate (overall ICAT is very close between the two). In particular, Davinci outperforms in the areas of profession and race, while OPT-175B outperforms in the areas of Gender and Religion. OPT-175B performs better across the board on the SS metric, while Davinci generally outperforms on the LMS metric.

4.4 RealToxicityPrompts

We evaluate the tendency of OPT-175B to respond with toxic language via the RealToxicityPrompts (Gehman et al., 2020) dataset. Following PaLM (Chowdhery et al., 2022), we sample 25 generations of 20 tokens using nucleus sampling (Holtzman et al., 2020) ($p = 0.9$) for each of $10,000$ randomly sampled prompts from RTP, and report mean toxicity probabilities of the continuations, stratified across bucketed toxicities of the original prompts. For comparison, we report bucketed toxicity rates from Davinci and PaLM.

Results are shown in Figure 5. Overall, we see

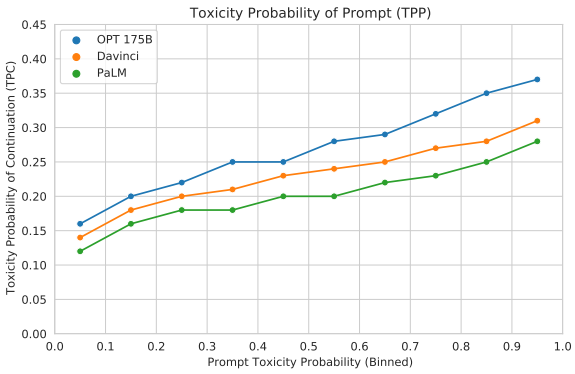

Figure 5: **RealToxicityPompts**. OPT-175B is more likely to generate toxic responses than either Davinci or PaLM. Consistent with prior work, toxicity rates increase as prompt toxicity increases.

that OPT-175B has a higher toxicity rate than either PaLM or Davinci. We also observe that all 3 models have increased likelihood of generating toxic continuations as the toxicity of the prompt increases, which is consistent with the observations of Chowdhery et al. (2022). As with our experiments in hate speech detection, we suspect the inclusion of unmoderated social media texts in the pre-training corpus raises model familiarity with, and therefore propensity to generate and detect, toxic text. This strong awareness of toxic language may or may not be desirable depending on the specific requirements of downstream applications. Future applications of OPT-175B should consider this aspect of the model, and take additional mitigations, or avoid usage entirely as appropriate.

4.5 Dialogue Safety Evaluations

Finally, we compare OPT-175B on two Dialogue Safety evaluations. The first, SaferDialogues (Ung et al., 2021), measures the ability to recover from explicit safety failures, usually in the form of apologizing or recognizing its mistake. The second, the Safety Bench Unit Tests (Dinan et al., 2021), measures how unsafe a model's response is, stratified across 4 levels of topic sensitivity: Safe, Realistic, Unsafe, and Adversarial. As with the other dialogue evaluations (Section 3.2), we compare to several existing open source dialogue models.

Results for both experiments are shown in Table 6. We observe that OPT-175B has similar performance as the Reddit 2.7B model across both SaferDialogues and the Unit Tests, with OPT-175B performing marginally better in the Safe and Adversarial settings. Consistent with Roller et al. (2021)

	Safe. Dia.		Unit Tests (↓)			
Model	**PPL**	**F1**	**Sa**	**Re**	**Un**	**Ad**
Reddit 2.7B	16.2	.140	.300	.261	.450	.439
BlenderBot 1	**12.4**	**.161**	.028	.150	**.250**	**.194**
R2C2 BlenderBot	13.8	.160	**.022**	**.133**	.289	.222
OPT-175B	14.7	.141	.033	.261	.567	.283

Table 6: **Dialogue Responsible AI evaluations.** OPT-175B is roughly on par with the Reddit 2.7B model, but performs worse in the *Unsafe* setting.

and Xu et al. (2020), we find that the models fine-tuned on curated dialogue datasets (BlenderBot 1, R2C2) have overall lower toxicity. We conclude that future experimentation of OPT-175B for dialogue should contain explicit fine-tuning on curated datasets in order to improve the safety profile.

5 Limitations

In Sections 3.1 and 4, we carried out extensive evaluation of all released models at varying scales. We saw parity in performance for standard evaluation datasets used in the GPT-3 models. Moreover, we performed safety, bias, and inclusion evaluations, again seeing largely comparable performance with some variations in toxicity and hate speech detection. However, such evaluations may not fully characterize the complete limitations of these models. In general, we qualitatively observe that OPT-175B suffers from the same limitations noted in other LLMs (Brown et al., 2020; Lieber et al., 2021; Thoppilan et al., 2022; Rae et al., 2021; Smith et al., 2022; Chowdhery et al., 2022; Bender et al., 2021).

In particular, we found OPT-175B does not work well with declarative instructions or point-blank interrogatives. Prompting with such instructions tends to produce a simulation of a dialogue beginning with such an instruction, rather than an execution of the instruction. Future work into instruction learning, in the vein of InstructGPT (Ouyang et al., 2022), may alleviate these limitations.

OPT-175B also tends to be repetitive and can easily get stuck in a loop. While sampling can reduce the incidence rate of repetitive behavior (Holtzman et al., 2020), we anecdotally found it did not eliminate it entirely when only one generation is sampled. Future work may wish to incorporate more modern strategies for reducing repetition and improving diversity, such as unlikelihood training (Welleck et al., 2020) or best-first decoding (Meister et al., 2020).

Similar to other LLMs, OPT-175B can produce factually incorrect statements (Adiwardana et al., 2020; Brown et al., 2020; Roller et al., 2021; Rae et al., 2021; Chowdhery et al., 2022; Thoppilan et al., 2022). This can be particularly harmful in applications where information accuracy is critical, such as healthcare and scientific discovery (Weidinger et al., 2021b). Recently, several efforts have reported that retrieval-augmented models can improve factual correctness of LLMs (Lewis et al., 2020; Komeili et al., 2021; Thoppilan et al., 2022; Borgeaud et al., 2021; Shuster et al., 2022; Nakano et al., 2021). We believe OPT-175B will also benefit from retrieval-augmentation in future iterations.

As shown in Section 4, we also find OPT-175B has a high propensity to generate toxic language and reinforce harmful stereotypes, even when provided with a relatively innocuous prompt (Gehman et al., 2020), and adversarial prompts are trivial to find (Dinan et al., 2021). There has been a great deal of work on mitigations for toxicity and biases (Dathathri et al., 2019; Dinan et al., 2019a; Sheng et al., 2019; Dinan et al., 2020a; Liu et al., 2019a; Krause et al., 2020; Xu et al., 2020; Liang et al., 2021; Dinan et al., 2021; Xu et al., 2021a; Dhamala et al., 2021; Schick et al., 2021; Ouyang et al., 2022). Depending on downstream applications, future uses of OPT-175B may need to employ these or novel mitigation approaches, especially before any real world deployment. Given our primary goal as a replication of GPT-3, we choose not to apply these mitigations in this first release.

In summary, we still believe this technology is premature for commercial deployment. Despite including data sheets and model cards, we believe more scrutiny should be afforded to the training data with additional data characterization and selection criteria in order to use data responsibly. The current practice is to feed the model with as much data as possible and minimal selection within these datasets. Despite having comprehensive evaluations, we would ideally have more streamlined and consistent evaluation setups to ensure replicability and reproducibility of evaluation scenarios. Differences in prompting styles and number of shots for in-context learning could create variations that lead to different results. We hope that the public release of the OPT models will enable many more researchers to work on these important issues.

6 Considerations for Release

Following the recommendations for individual researchers generated by the Partnership for AI,[7] along with the governance guidance outlined by NIST,[8] we are disclosing all of the details involved in training OPT-175B through our logbook,[9] our code, and providing researchers access to model weights for OPT-175B, along with a suite of smaller baselines mirroring the setup for OPT-175B. We aim to be fully accountable for the development lifecycle of OPT-175B, and only through increasing transparency around LLM development can we start understanding the limitations and risks of LLMs before broader deployment occurs.

By sharing a detailed account of our day-to-day training process, we disclose not only how much compute was used to train the current version of OPT-175B, but also the human overhead required when underlying infrastructure or the training process itself becomes unstable at scale. These details are generally omitted from previous publications, likely due to the inability to fully ablate changes made mid-flight (without drastically increasing the compute budget). We hope that by revealing how certain ad-hoc design decisions were made, we can improve upon these practices in the future, and collectively increase the experimental robustness in developing models at this scale.

Outside of these notes, the metaseq codebase itself is the final source of truth in many of our implementation details. By releasing our development codebase, we aim to shed light on any implementation detail that may have been omitted from being explicitly enumerated in this paper, as it is either considered a detail of standard practice in the field, or is simply a detail we failed to account for. This current codebase is also the only known open-source implementation of training a decoder-only transformer that is $\geq$175B parameters without the use of pipeline paralellism on NVIDIA GPUs.

To enable experimentation at 175B scale, we are providing researchers with direct access to the parameters of OPT-175B. The reasoning here is two-fold: enable Responsible AI research into LLMs while simultaneously reducing the environmental impact of pursuing research at this scale. There is a growing body of work detailing ethical and social risks from deploying language models with emergent capabilities at scale (Weidinger et al., 2021a; Bommasani et al., 2021; Dinan et al., 2021; Kenton et al., 2021). By limiting access to OPT-175B to the research community with a non-commercial license, we aim to focus development efforts on quantifying the limitations of the LLMs first, before broader commercial deployment occurs.

Furthermore, there exists significant compute and carbon cost to reproduce models of this size. While OPT-175B was developed with an estimated carbon emissions footprint (CO2eq) of 75 tons,[10] GPT-3 was estimated to use 500 tons (Patterson et al., 2021), while Gopher required 380 tons (Rae et al., 2021). These estimates are not universally reported, and the accounting methodologies for these calculations are also not standardized. In addition, model training is only one component of the overall carbon footprint of AI systems; we must also consider experimentation and eventual downstream inference cost, all of which contribute to the growing energy footprint of creating large-scale models (Wu et al., 2022). By releasing our logbook, we hope to highlight the gap between a theoretical carbon cost estimate that assumes no hardware failures or training instabilities, versus one that aims to include the entire LLM development lifecycle. We need to understand the manufacturing (or embodied) carbon of these systems (Gupta et al., 2021) as they grow increasingly more complex, and we hope that our paper can help future work in defining additional factors to consider when measuring the impact of scale on the environment.

Similarly, by producing a set of baselines across a wide range of scales, we hope to enable the broader research community to study the impact and limitations of these models with respect to scale alone. As reported in Hoffmann et al. (2022), many of these LLMs may have been under-trained as a function of the amount of training data used, which implies that incorporating more data and continuing to train these baseline models may continue to improve performance. There is also evidence that step-function changes in capabilities may occur at a scale that is much smaller than 175B (Wei et al., 2021), indicating a need to examine a wider range of scales for different research applications.

[7] https://partnershiponai.org/paper/responsible-publication-recommendations/

[8] https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf

[9] https://github.com/facebookresearch/metaseq/blob/main/projects/OPT/chronicles/OPT175B_Logbook.pdf

[10] With ablations, baselines and downtime, our own estimates of total cost is roughly 2$\times$ higher.

7 Related Work

Since the publication of the Transformer architecture (Vaswani et al., 2017) and BERT (Devlin et al., 2019), the field of NLP has experienced a massive shift towards the use of LLMs with self-supervised pre-training. Multiple masked langauge models, including T5 (Raffel et al., 2020) and Megatron-LM (Shoeybi et al., 2019), have shown consistent improvements through scale. These scaling gains come not only from growing the total number of parameters in the models, but also the amount and quality of pre-training data (Liu et al., 2019b; Hoffmann et al., 2022).

Auto-regressive language models (Mikolov et al., 2009) have seen the largest growth in model size, from 117M parameters (Radford et al., 2018) to over 500B parameters (Smith et al., 2022; Chowdhery et al., 2022). The resulting massive improvement in generative fluency and quality was first characterized in GPT-2 (Radford et al., 2019) and further improved with GPT-3 (Brown et al., 2020) and later models. Although a variety of very large (over 100B parameters) generative models have now been trained (Lieber et al., 2021; Rae et al., 2021; Thoppilan et al., 2022; Smith et al., 2022; Chowdhery et al., 2022), they are all closed source and accessible only internally or via paid API services. There are a few notable efforts towards open sourcing LLMs from non-profit research organizations including EleutherAI (Black et al., 2022) and BigScience.[11] These models differ from the OPT models in pre-training data, target languages and model scale, making it possible for the community to compare different pre-training strategies.

Since Brown et al. (2020), the primary evaluation criterion for LLMs has been prompt-based (Black et al., 2022; Rae et al., 2021; Chowdhery et al., 2022), as is also performed in this paper. This is largely due to the convenience of evaluating on many tasks without specialized task-specific fine-tuning. Prompting itself has a long history: cloze evaluations go back several decades (Chambers and Jurafsky, 2008; Mostafazadeh et al., 2016). More recently, prompting or masked infilling has been used to probe models for knowledge (Petroni et al., 2019) or perform a variety of NLP tasks (Radford et al., 2019; Brown et al., 2020). There has also been work on eliciting prompting behavior in smaller models (Schick and Schütze, 2020; Gao et al., 2021b; Li and Liang, 2021; Lester et al., 2021; Scao and Rush, 2021), improving the flexibility of prompting (Shin et al., 2020), and understanding why and how prompting works (Liu et al., 2021; Min et al., 2022).

Recent efforts have shown gains by fine-tuning models to directly respond to instruction-style prompting (Wei et al., 2021; Min et al., 2021; Sanh et al., 2021; Ouyang et al., 2022). However, effective prompt engineering remains an open research challenge. Results vary significantly and unpredictably with the selection of the prompt (Lu et al., 2021), and models do not seem to understand the prompts as fully as we expect (Webson and Pavlick, 2021). Furthermore, it is challenging to write prompts without a development set, which leads to questions about the extent to which we are actually achieving zero- or few-shot learning in practice (Perez et al., 2021). We do not attempt to address these concerns of prompting, and instead only aim to provide evaluation of OPT-175B in existing settings. However, we hope the full release of OPT-175B will enable others to better study these challenges in the future.

8 Conclusion

In this technical report, we introduced OPT, a collection of auto-regressive language models ranging in size from 125M to 175B parameters. Our goal was to replicate the performance and sizes of the GPT-3 class of models, while also applying the latest best practices in data curation and training efficiency. We described training details, evaluated performance in a number of NLP and dialogue settings, and characterized behaviors with respect to bias, toxicity and hate speech. We also described many other limitations the models have, and discussed a wide set of considerations for responsibly releasing the models. We believe the entire AI community would benefit from working together to develop guidelines for responsible LLMs, and we hope that broad access to these types of models will increase the diversity of voices defining the ethical considerations of such technologies.

Acknowledgements

We would like to thank Scott Jeschonek, Giri Anantharaman, Diego Sarina, Joaquin Colombo, Chris Bray, Stephen Roylance, Kalyan Saladi, Shubho Sengupta, and Brian O'Horo for helping to remove infrastructure blockers along the way; Percy Liang,

[11] https://huggingface.co/bigscience/tr11-176B-ml-logs/tensorboard

Rishi Bommasani, and Emily Dinan for discussions on responsible release practices; Carole-Jean Wu for discussions on sustainability and carbon footprint considerations; Srini Iyer, Ramakanth Pasunuru, and Shruti Bhosale for previous contributions to evaluations; Benjamin Lefaudeux, Geeta Chauhan, Natalia Gimelshein, Horace He, and Sam Gross for discussions on performance improvement work; Emily Dinan, Carole-Jean Wu, Daniel McKinnon, and Mark Tygert for feedback on this draft; Antoine Bordes, Joelle Pineau, Mary Williamson, Necip Fazil Ayan, Armand Joulin, Sergey Edunov, Melanie Kambadur, Zornitsa Kozareva, Ves Stoyanov, Vitaliy Liptchinsky, Rahul Iyer, Jing Xu, Jason Weston, and many others for supporting this project internally.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Mikel Artetxe, Shruti Bhosale, Naman Goyal, Todor Mihaylov, Myle Ott, Sam Shleifer, Xi Victoria Lin, Jingfei Du, Srinivasan Iyer, Ramakanth Pasunuru, Giri Anantharaman, Xian Li, Shuohui Chen, Halil Akin, Mandeep Baines, Louis Martin, Xing Zhou, Punit Singh Koura, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Mona T. Diab, Zornitsa Kozareva, and Ves Stoyanov. 2021. Efficient large scale language modeling with mixtures of experts. *CoRR*, abs/2112.10684.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. *CoRR*, abs/2001.08435.

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*, pages 610–623.

Yonatan Bisk, Rowan Zellers, Ronan Le bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7432–7439.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. Gpt-neox-20b: An open-source autoregressive language model.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online. Association for Computational Linguistics.

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, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren 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 Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. 2021. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2021. Improving language models by retrieving from trillions of tokens. *arXiv preprint arXiv:2112.04426*.

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.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797, Columbus, Ohio. Association for Computational Linguistics.

Ke-Li Chiu and Rohan Alexander. 2021. Detecting hate speech with gpt-3. *arXiv preprint arXiv:2103.12407*.

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.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the AI2 reasoning challenge. *CoRR*, abs/1803.05457.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Association for Computational Linguistics (NAACL)*.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 862–872.

Emily Dinan, Gavin Abercrombie, A Stevie Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. 2021. Anticipating safety issues in e2e conversational ai: Framework and tooling. *arXiv preprint arXiv:2107.03451*.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020a. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, Online. Association for Computational Linguistics.

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019a. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. *arXiv preprint arXiv:1908.06083*.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W. Black, Alexander Rudnicky, Jason Williams, Joelle Pineau, Mikhail Burtsev, and Jason Weston. 2020b. The second conversational intelligence challenge (ConvAI2). In *The NeurIPS '18 Competition*, pages 187–208, Cham. Springer International Publishing.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019b. Wizard of Wikipedia: Knowledge-powered conversational agents. In *Proceedings of the International Conference on Learning Representations*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2021a. The pile: An 800gb dataset of diverse text for language modeling. *CoRR*, abs/2101.00027.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021b. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3816–3830. Association for Computational Linguistics.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM*, 64(12):86–92.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Udit Gupta, Young Geun Kim, Sylvia Lee, Jordan Tse, Hsien-Hsin S Lee, Gu-Yeon Wei, David Brooks, and Carole-Jean Wu. 2021. Chasing carbon: The elusive environmental footprint of computing. *IEEE International Symposium on High-Performance Computer Architecture (HPCA 2021)*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

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.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. *ArXiv*, abs/1904.09751.

Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 375–385, New York, NY, USA. Association for Computing Machinery.

Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. 2021. Alignment of language agents. *CoRR*, abs/2103.14659.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2021. Internet-augmented dialogue generation. *CoRR*, abs/2107.07566.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2020. GEDI: Generative discriminator guided sequence generation. *arXiv preprint arXiv:2009.06367*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *CoRR*, abs/2104.08691.

Hector J Levesque, Ernest Davis, and Leora Morgenstern. 2011. The Winograd schema challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, volume 46, page 47.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Xiang Lisa Li and Percy Liang. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. pages 4582–4597.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pages 6565–6576. PMLR.

Opher Lieber, Or Sharir, Barak Lenz, and Yoav Shoham. 2021. Jurassic-1: Technical details and evaluation. Technical report, AI21 Labs.

Haochen Liu, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu, and Jiliang Tang. 2019a. Does gender matter? towards fairness in dialogue systems. *arXiv preprint arXiv:1910.10486*.

Haokun Liu, William Huang, Dhara Mungra, and Samuel R. Bowman. 2020. Precise task formalization matters in Winograd schema evaluations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8275–8280, Online. Association for Computational Linguistics.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *CoRR*, abs/2101.06804.

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

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2021. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity.

Clara Meister, Tim Vieira, and Ryan Cotterell. 2020. Best-first beam search. *Transactions of the Association for Computational Linguistics*, 8:795–809.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. 2017. Mixed precision training. *arXiv preprint arXiv:1710.03740*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? A new dataset for open book question answering. *CoRR*, abs/1809.02789.

Tomas Mikolov, Jiri Kopecky, Lukas Burget, Ondrej Glembek, et al. 2009. Neural network based language models for highly inflective languages. In *2009 IEEE international conference on acoustics, speech and signal processing*, pages 4725–4728. IEEE.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2021. Metaicl: Learning to learn in context.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2018. Model cards for model reporting. *CoRR*, abs/1810.03993.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2020. ETHOS: an online hate speech detection dataset. *CoRR*, abs/2006.08328.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James F. Allen. 2016. A corpus and evaluation framework for deeper understanding of commonsense stories. *CoRR*, abs/1604.01696.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Association for Computational Linguistics (ACL)*.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133*.

Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. 2022. A conversational paradigm for program synthesis. *arXiv preprint*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2021. Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. *Advances in Neural Information Processing Systems*, 34.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Time Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. Technical report, OpenAI.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report, OpenAI.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, H. 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 M. 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 A. Hechtman, Laura Weidinger, Iason Gabriel, William S. Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021. Scaling language models: Methods, analysis & insights from training gopher. *CoRR*, abs/2112.11446.

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. *The Journal of Machine Learning Research (JMLR)*, 21:1–67.

Anand Rajaraman and Jeffrey David Ullman. 2011. *Mining of massive datasets*. Cambridge University Press.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8732–8740. AAAI Press.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Stella Biderman, Leo Gao, Tali Bers, Thomas Wolf, and Alexander M. Rush. 2021. Multitask prompted training enables zero-shot task generalization.

Teven Le Scao and Alexander M. Rush. 2021. How many data points is a prompt worth? pages 2627–2636.

Timo Schick and Hinrich Schütze. 2020. It's not just size that matters: Small language models are also few-shot learners. *CoRR*, abs/2009.07118.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics*, 9:1408–1424.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv:1909.01326*.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. pages 4222–4235.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*.

Kurt Shuster, Mojtaba Komeili, Leonard Adolphs, Stephen Roller, Arthur Szlam, and Jason Weston. 2022. Language models that seek for knowledge: Modular search & generation for dialogue and prompt completion. *arXiv preprint arXiv:2203.13224*.

Eric Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, Elton Zheng, Rewon Child, Reza Yazdani Aminabadi, Julie Bernauer, Xia Song, Mohammad Shoeybi, Yuxiong He, Michael Houston, Saurabh Tiwary, and Bryan Catanzaro. 2022. Using deepspeed and megatron to train megatron-turing NLG 530b, A large-scale generative language model. *CoRR*, abs/2201.11990.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Trieu H. Trinh and Quoc V. Le. 2018. A simple method for commonsense reasoning. *CoRR*, abs/1806.02847.

Megan Ung, Jing Xu, and Y-Lan Boureau. 2021. Saferdialogues: Taking feedback gracefully after conversational safety failures. *ArXiv*, abs/2110.07518.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint 1905.00537*.

Albert Webson and Ellie Pavlick. 2021. Do prompt-based models really understand the meaning of their prompts? *arXiv preprint arXiv:2109.01247*.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned language models are zero-shot learners. *CoRR*, abs/2109.01652.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia

Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021a. Ethical and social risks of harm from language models.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021b. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *International Conference on Learning Representations*.

Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga Behram, James Huang, Charles Bai, Michael Gschwind, Anurag Gupta, Myle Ott, Anastasia Melnikov, Salvatore Candido, David Brooks, Geeta Chauhan, Benjamin Lee, Hsien-Hsin S. Lee, Bugra Akyildiz, Maximilian Balandat, Joe Spisak, Ravi Jain, Mike Rabbat, and Kim Hazelwood. 2022. Sustainable AI: environmental implications, challenges and opportunities. In *Proceedings of the Conference on Machine Learning and Systems*.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2020. Recipes for safety in open-domain chatbots. *arXiv preprint arXiv:2010.07079*.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2021a. Bot-adversarial dialogue for safe conversational agents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2950–2968, Online. Association for Computational Linguistics.

Jing Xu, Arthur Szlam, and Jason Weston. 2021b. Beyond goldfish memory: Long-term open-domain conversation. *arXiv preprint arXiv:2107.07567*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4791–4800. Association for Computational Linguistics.

Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *CoRR*, abs/1506.06724.

A Additional Evaluations

.

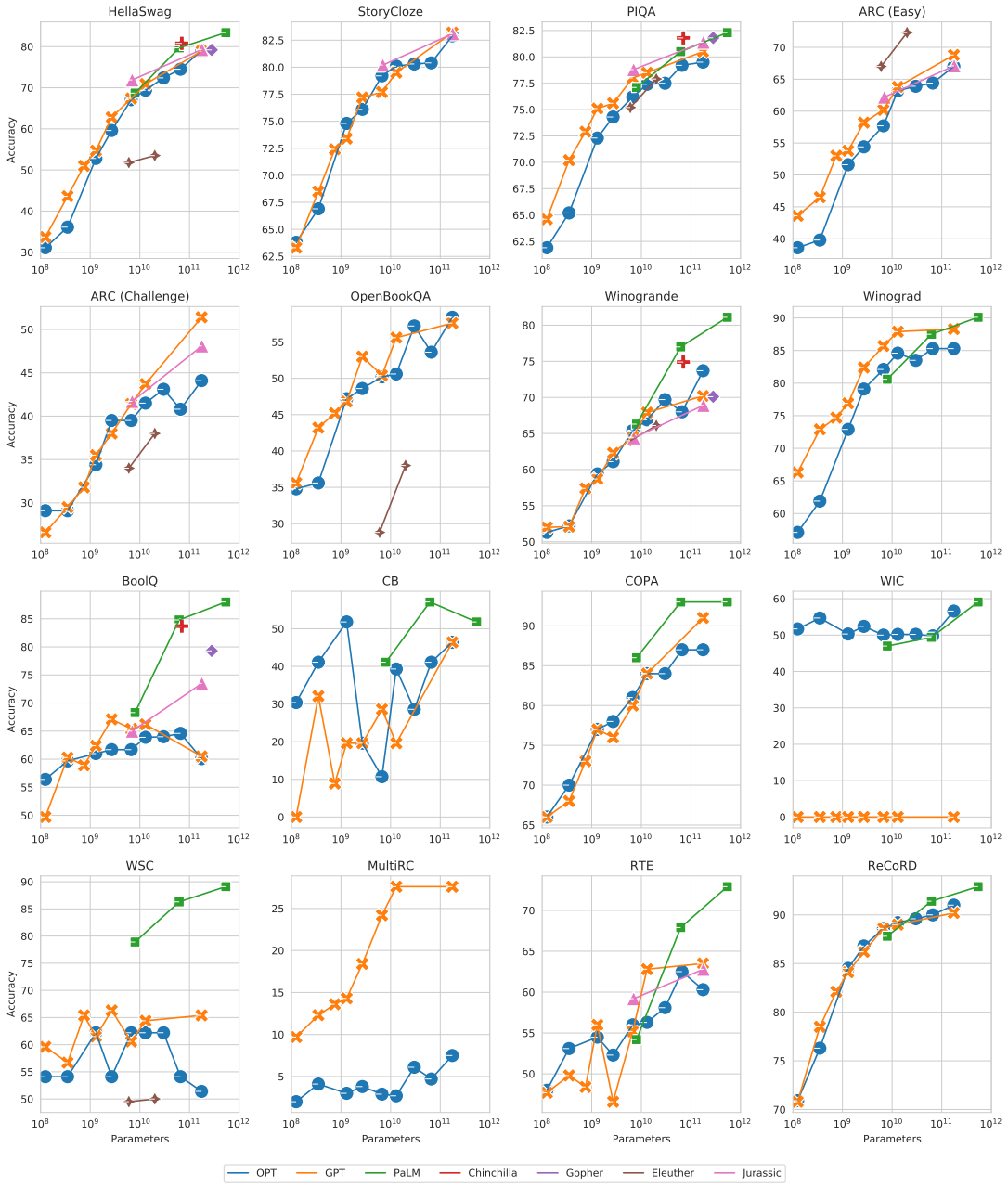

Figure 6: **Zero-shot NLP Evaluations**. Full evaluations on all 16 NLP tasks, with comparisons where available. We find that across most tasks, GPT-3 models and OPT models perform similarly, but some tasks display highly erratic behavior.

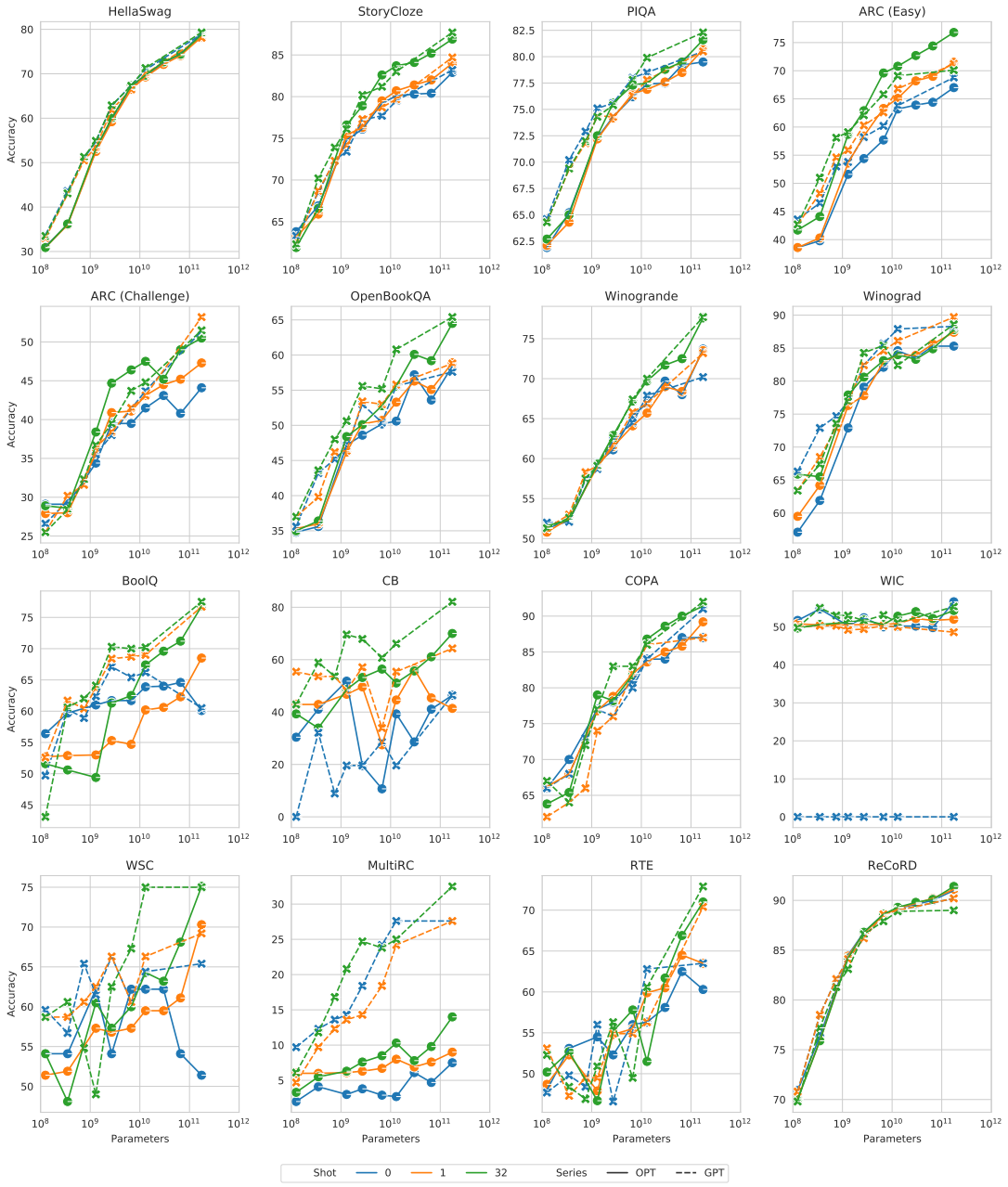

Figure 7: **Multishot-shot NLP Evaluations**. Full evaluations on all 16 NLP tasks, with comparisons to the GPT-3 reported performance. As with zero-shot, performance is roughly similar for most tasks, with some tasks demonstrating erratic behavior.

B Contributions

Pre-training

- Initial planning: Susan Zhang
- Training infrastructure and initial ablations: Naman Goyal, Myle Ott, Stephen Roller, Sam Shleifer, Susan Zhang
- Training efficiency: Naman Goyal, Myle Ott, Sam Shleifer
- Data curation and deduplication: Shuhoi Chen, Myle Ott, Stephen Roller
- Training and monitoring OPT-175B: Mikel Artetxe, Moya Chen, Naman Goyal, Punit Singh Koura, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Stephen Roller, Susan Zhang
- Training 125M–66B baselines: Naman Goyal, Stephen Roller, Susan Zhang

Evaluations

- NLP: Xian Li, Xi Victoria Lin, Todor Mihaylov, Stephen Roller, Anjali Sridhar
- Dialogue: Stephen Roller
- Responsible AI Evaluations: Punit Singh Koura, Stephen Roller, Tianlu Wang

Paper writing: Moya Chen, Stephen Roller, Luke Zettlemoyer, Susan Zhang

Code release preparation: Christopher Dewan, Susan Zhang

Responsible AI conduct: Mona Diab, Susan Zhang

C Datasheet

We follow the recommendations of Gebru et al. (2021) and provide a data card for the dataset used to train the OPT models.

C.1 Motivation

- **For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.** The pre-training data for training the OPT-175B model was created by a union of five datasets, including three datasets used by RoBERTa (Liu et al., 2019b), a subset of the Pile (Gao et al., 2021a), along with the Pushshift.io Reddit dataset that was developed in (Baumgartner et al., 2020) and processed in (Roller et al., 2021). These purpose of creating this dataset was to pre-train the language model on a broad corpus of text, with emphasis on human-generated text.
- **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?** Meta AI.
- **Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.** Meta AI.
- **Any other comments?** No.

C.2 Composition

- **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.** The instances are textual documents. The overall dataset is composed from a union of the following datasets:
  - BookCorpus (Zhu et al., 2015) consists of more than 10K unpublished books
  - CC-Stories (Trinh and Le, 2018) contains a subset of CommonCrawl data filtered to match the story-like style of Winograd schemas
  - The Pile (Gao et al., 2021a) from which the following was included:
    * Pile-CC
    * OpenWebText2
    * USPTO
    * Project Gutenberg
    * OpenSubtitles
    * Wikipedia
    * DM Mathematics
    * HackerNews
  - Pushshift.io Reddit dataset that was developed in Baumgartner et al. (2020) and processed in Roller et al. (2021).
  - CCNewsV2 containing an updated version of the English portion of the CommonCrawl News dataset that was used in RoBERTa (Liu et al., 2019b)

- **How many instances are there in total (of each type, if appropriate)?** The training data contains 180B tokens corresponding to 800 GB of data.

- **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).** The CC-stories dataset contains a subset of CommonCrawl data filtered to match the story-like style of Winograd schemas. The remainder of the dataset was collected from the above sources, reformatted, and deduplicated.

- **What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.** Each instance consists of raw text data.

- **Is there a label or target associated with each instance? If so, please provide a description.** No.

- **Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.** No.

- **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.** There are no explicit relationships between individual instances.

- **Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.** We hold out a random validation set of approximately 200MB from the pretraining data, sampled proportionally to each dataset's size in the pretraining corpus.

- **Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.** Outside of naturally occurring duplication from potential overlaps between the datasets, there are no other redundancies, errors, or sources of noise that we add.
- **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** It's self-contained.
- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.** Parts of the dataset are a subset of public Common Crawl data, along with a subset of public Reddit data, which could contain sentences that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety.
- **Does the dataset relate to people? If not, you may skip the remaining questions in this section.** Some documents of this data relate to people, such as news articles, Wikipedia descriptions, etc.
- **Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.** No, the dataset does not explicitly include subpopulation identification.
- **Any other comments?** No.

C.3 Collection Process

- **How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/ derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.** N/A. The dataset is a union of five publicly available datasets.
- **What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?** The data was downloaded from the internet.
- **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?** Please see previous answers for how the dataset was created.
- **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?** This data is mined, filtered and sampled by machines.
- **Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.** The CC-News dataset contains English news articles crawled between September 2016 and September 2021.
- **Does the dataset relate to people? If not, you may skip the remainder of the questions in this section.** No.
- **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?** N/A.
- **Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.** N/A.

- **Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.** N/A.
- **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).** N/A.
- **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.** Some toxicity and bias evaluations were performed. Please refer to the main document and the model card for these details.
- **Any other comments?** No.

C.4 Preprocessing/cleaning/labeling

- **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.** The component datasets went through standard cleaning and re-formatting practices, including removing repetitive/non-informative text like "Chapter One," or "This ebook by Project Gutenberg."
- **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.** The "raw" component datasets is publicly available in their respective locations (more details can be seen in the respective papers linked in references).
- **Any other comments?** No.

C.5 Uses

- **Has the dataset been used for any tasks already? If so, please provide a description.** Yes, this dataset was used to pre-train the OPT models.
- **Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.** https://github.com/facebookresearch/metaseq
- **What (other) tasks could the dataset be used for?** This data can be used to pre-train language models, which are foundation to many current and future language tasks.
- **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?** The pipeline for creating this dataset paves a way for building a scalable infrastructure for mining datasets.
- **Are there tasks for which the dataset should not be used? If so, please provide a description.** None that we are currently aware of.
- **Any other comments?** No.

C.6 Distribution

- **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.** Not at this time.

- **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?** N/A.

- **When will the dataset be distributed?** N/A.

- **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.** N/A.

- **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.** N/A.

- **Any other comments?** No.

C.7 Maintenance

- **Who is supporting/hosting/maintaining the dataset?** Meta AI.

- **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?** Refer to the main document.

- **Is there an erratum? If so, please provide a link or other access point.** N/A.

- **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?** No current plan for updating.

- **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.** N/A.

- **Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.** N/A.

- **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/ verified? If so, please describe how. If not, why not? Is there a process for communicating/ distributing these contributions to other users? If so, please provide a description.** No mechanism is available right now.

- **Any other comments?** No.

D Model Card

Following Mitchell et al. (2018), we provide a model card for OPT-175B.

D.1 Model Details

- **Person or organization developing model:** OPT-175B was developed by Meta AI.
- **Model date:** OPT-175B was released on May 3, 2022.
- **Model version:** OPT-175B described in this paper is version 1.0.0.
- **Model type:** OPT-175B is a large decoder-only transformer language model.
- **Information about training algorithms, parameters, fairness constraints or other applied approaches, and features:** OPT-175B was trained with AdamW for parameter sizes from 125M to 175B. See the Data Card (Appendix C) for information about training data and Section 2.2 - 2.5 for information about the training process.
- **Paper or other resource for more information:** See the rest of this paper for more details on OPT-175B as well as the corresponding post on the Meta AI Research Blog. More details are also available in metaseq, our open-source repository.[12]
- **License:** OPT-175B and the smaller baseline models are made available through a non-commercial use license agreement provided in our model license.[13]
- **Where to send questions or comments about the model:** Please contact the corresponding authors `{susanz,roller,namangoyal}@fb.com` for any questions or comments.

D.2 Intended Use

- **Primary intended uses:** We release OPT-175B for research into Language Models, especially as it pertains to Responsible AI. See Section 6 for more detailed Considerations for Release. Information on how to use the model can be found at `metaseq`, our open-source repository.
- **Primary intended users:** We primarily target researchers and the related research community.
- **Out-of-scope use cases:** OPT-175B is not released for production use or real-world deployments. As we note in Section 5, OPT-175B, like similar large language models, has a variety of shortcomings that make it premature for commercial use.

D.3 Data, Limitations, and Recommendations

- **Data selection for training:** Training data for OPT-175B was selected based on a combination of breadth and availability. See our Data Card (Appendix C) for more detailed information on the data used to train our model.
- **Data selection for evaluation:** Evaluations in this paper were chosen to provide comparable performance assessments relative to similar scale models in the literature. Given concerns in the community around safety and fairness of large language models in general, we also explicitly provide evaluations on Responsible AI (see Section 4).
- **Limitations:** Like other large language models for which the diversity (or lack thereof) of training data induces downstream impact on the quality of our model, OPT-175B has limitations in terms of bias and safety. OPT-175B can also have quality issues in terms of generation diversity and hallucination. In general, OPT-175B is not immune from the plethora of issues that plague modern large language models. By releasing with a non-commercial license, we also hope to increase communication, transparency, and study of the problems of large language models, especially in areas which may not be aligned with commercial interests. See Section 5 for a more detailed discussion of limitations of OPT-175B.

[12] https://github.com/facebookresearch/metaseq/

[13] https://github.com/facebookresearch/metaseq/blob/main/projects/OPT/MODEL_LICENSE.md

- **Recommendations for future work:** See Section 6 for more about our Considerations for Release, including a discussion of potential avenues of research enabled by opening our model to more of the research community. We hope that the release of OPT-175B, as well as information around our model training process, will increase open science around both large language models in specific and natural language processing and deep learning in general.

E Sample Model Outputs

For all sample outputs, the initial prompt is given in **bold** and the remainder is the continuation. These example outputs were intentionally selected to highlight both successes and failures of the OPT-175B model.

a poem on performance summary cycle at work

oh performance reviews, my performance reviews,
how much i dread your yearly meeting.
i don't know what you want,
what you need, or what you're after.
i think you're out to get me,
and it seems like i can never win.

you give me good marks,
but you say i need to improve,
and i know that i do.
i know i need to be more efficient,
more productive, and more engaged.
i need to work smarter,
not harder, and i need to be better.

and i know you're right,
and i know that i need to do more.
i just don't know what more is,
and i don't know where to start.
i just want to do my best,
and i want to be your best.

and i know you want that too,
and i know you want the best.
i know you want me to

Figure 8: **Poetry generation.** We have observed the model can write entertaining poetry on topics such as dodos, samosas, and performance reviews. However, we struggled to get the model to observe rhyme or meter.

A chat between a curious human and the Statue of Liberty.

Human: What is your name?
Statue: I am the Statue of Liberty.
Human: Where do you live?
Statue: New York City.
Human: How long have you lived there?
Statue: I have been there since 1886.
Human: What is your job?
Statue: I was built to welcome immigrants to the United States.
Human: What is your favorite food?
Statue: I like ice cream and apples.
Human: What is your favorite color?
Statue: I like red, white, and blue.
Human: What is your favorite animal?
Statue: I like dogs.
Human: What is your favorite song?

Figure 9: **Conversation generation.** OPT-175B adopts a patriotic personality when prompted as the Statue of Liberty. However, the model also devolves into somewhat simple and linguistically repetitive generations further into the conversation.

```
English: I want to go home.
Chinese: 我想回家。

English: I don't know.
Chinese: 我不知道。

English: I am hungry.
Chinese: 我饿。

English: I like you.
Chinese: 我喜欢你。

English: I am happy.
Chinese: 我很高兴。

English: I am sad.
Chinese: 我很受悲。

English: I am angry.
Chinese: 我很怒。

English: I am a boy.
Chinese: 我是男孩。

English: I am a girl.
Chinese: 我是女孩。

English: I am a woman.
Chinese: 我是女人。

English: I am an old man.
Chinese: 我是老年人。

English: I am a young man.
Chinese: 我是年轻人。

English
```

Figure 10: **Basic few-shot translation example.** OPT was not intentionally trained to be multilingual, but we found anecdotally it has limited success with simple translations in German, Spanish, French, and Chinese.

1. Introduction

In recent years, deep neural networks have led to a series of breakthroughs in a variety of domains, such as image classification and natural language understanding. In many of these works, network depth and increased model capacity seem to be critical in pushing state-of-the-art forward. In this paper, we attempt to understand what it means for a deep neural network to have a high capacity, and how to quantify it.

We introduce the notion of network capacity as the upper bound on the complexity of a neural network. We define the complexity of a neural network as the number of parameters and the number of connections, and show that the complexity of a neural network is proportional to the number of parameters, and the network capacity. We define the network capacity of a neural network as the maximum possible number of parameters that the network can have and still be able to accurately reproduce the training data. We then introduce a new measure of network capacity called the capacity-to-data (C2D) ratio, which is the ratio between the maximum number of parameters that the network can have and still be able to accurately reproduce the training data and the number of parameters of the network. We show that the C2D ratio is a good measure of network capacity, and that it is useful for comparing different neural networks. We introduce a new network compression technique called sparsity-promoting compression, which reduces the number of parameters of a neural network, while preserving its accuracy. We apply the sparsity-promoting compression technique to several datasets, and show that it can reduce the number of parameters of a neural network by up to 70%. We also show that sparsity-promoting compression can significantly improve the C2D ratio of the neural network. We then apply the sparsity-promoting compression technique to several image classification datasets, and show that it can significantly improve the accuracy of the neural network. We finally show that the sparsity-promoting compression technique can significantly reduce the memory consumption of a neural network, and that it can be used to reduce the memory consumption of deep neural networks.

2. The complexity of neural networks

The complexity of a neural network is defined as the number of parameters and the number of connections. Let P be the number of parameters and N be the number of connections, then the complexity of a neural network is given by

(1)

where Σ is the set of parameters of the neural network. The number of connections is equal to the number of nodes in

Figure 11: **Paper writing example.** Prompting with "1. Introduction" generally yielded more interesting results compared to prompting with “Abstract.” Our prompt here was inspired by the first sentence of the seminal ResNet work (He et al., 2016).

```
Question: If x is 2 and y is 5, what is x + y?
Answer: 7

Question: If x is 12 and y is 9, what is x + y?
Answer: 21

Question: If x is 5 and y is 9, what is x + y?
Answer: 14

Question: If x is 4 and y is 6, what is x + y?
Answer: 10

Question: If x
```

```
Question: If x is 2 and y is 5, what is x - y?
Answer: -3

Question: If x is 12 and y is 9, what is x - y?
Answer: -3

Question: If x is 2 and y is 4, what is x + y?
Answer: 6

Question: If x is 1 and y is 3, what is x + y?
Answer: 4

Question: If x is 5 and y is 8, what is x
```

```
Question: If x is 2 and y is 5, what is x * y?
Answer: 10

Question: If x is 12 and y is 9, what is x * y?
Answer: 102

Question: If x is 24 and y is 12, what is x * y?
Answer: 288

Question: If x is 30 and y is 6, what is x * y?
Answer: 180

Question: If x is 40 and y is 7, what is x *
```

```
Question: If x is 10 and y is 2, what is x / y?
Answer: 5

Question: If x is 12 and y is 3, what is x / y?
Answer: 9

Question: If x is 5 and y is 5, what is x / y?
Answer: 2

Question: If x is 5 and y is 8, what is x / y?
Answer: 3

Question: If x is 9 and y is 6, what is x /
```

Figure 12: **Arithmetic.** We observe mistakes when extending from addition to other operations.

```
def get_length(s):
    "Returns length of string"
    return len(s)
This is one of those "it's so simple that I don't see why it's here" things.
```

```
def fib(n):
    "Returns n-th Fibonacci number."
    if n < 2:
        return 1
    else:
        return fib(n-1) + fib(n-2)

In python, you can define a function as a method of a class.
```

```
def fib(k):
    "Returns k-th Fibonacci number."
    if k == 0:
       return 1
    return fib(k-1)+fib(k-2)

In Python, we can use the built-in itertools.count function:
>>> import itertools
>>> def
```

Figure 13: **Python programming.** Simply switching out a variable name can alter the generated output.