GPT-NeoX-20B: An Open-Source Autoregressive Language Model

Anonymous ACL submission

Abstract

We introduce GPT-NeoX-20B, a 20 billion parameter autoregressive language model trained on the Pile, whose weights will be made freely and openly available to the public through a permissive license. It is, to the best of our knowledge, the largest dense autoregressive model that has publicly available weights at the time of submission. In this work, we describe GPT-NeoX-20B's architecture and training, and evaluate its performance. We open-source the training and evaluation code, as well as the model weights, at [URL will go here after peer review].

1 Introduction

Over the past several years, there has been an explosion in research surrounding large language models (LLMs) for natural language processing, catalyzed largely by the impressive performance of Transformer-based language models such as BERT (Devlin et al., 2019), GPT-2 (Radford et al., 2019), GPT-3 (Brown et al., 2020), and T5 (Raffel et al., 2020). One of the most impactful outcomes of this research has been the discovery that the performance of LLMs scales predictably as a power-law with the number of parameters, with architecture details such as width/depth ratio having a minimal impact on performance within a wide range (Kaplan et al., 2020). A consequence of this has been an abundance of research focusing on scaling Transformer models up to ever-larger scales, resulting in models of up to 530B parameters (Smith et al., 2022), a scale that would have been almost unthinkable just a few years prior.

Today, there are dozens of publicly acknowledged LLMs in existence. The largest have more than two orders of magnitude more parameters than GPT-2, and even at that scale there are nearly a dozen different models. However, these models are almost universally the protected intellectual property of large tech companies, and are gated behind a commercial API, available only upon request, or not available for outsider use at all. To our knowledge, the only freely and publicly available dense autoregressive language models larger than GPT-2 are GPT-Neo (2.7B parameters) (Black et al., 2021), CPM 1.0 (2.6B parameters) (Zhang et al., 2020), GPT-J-6B (Wang and Komatsuzaki, 2021), XGLM (7.5B parameters), Megatron-11B[1], CPM 2.1 (11B parameters) (Zhang et al., 2021), WuDao-GLM-XXL (11B parameters, with separate English and Chinese variants) (Du et al., 2021), Pangu-α-13B (Zeng et al., 2021), and the recently released FairSeq 6.7 and 13B (Artetxe et al., 2021) models.

In this paper, we release GPT-NeoX-20B, motivated by the belief that open access to LLMs is critical to advancing research in a wide range of areas—particularly in AI safety, mechanistic interpretability, and the study of how LLM capabilities scale. Many of the most interesting capabilities of LLMs appear to only emerge above a certain number of parameters, and such properties simply cannot be studied using smaller models. Although safety is often cited as a justification for keeping model weights private, we believe this self-imposed restriction is insufficient to prevent misuse, and moreover severely limits the probing and investigation of LLMs, especially for researchers not based at the small number of organizations that have access to state-of-the-art language models.

In the following sections, we give a broad overview of GPT-NeoX-20B's architecture and training hyperparameters, detail the hardware and software setup used for training and evaluation, and elaborate on the choices made when designing the training dataset and tokenization. We also address of some of the difficulties and unknowns we encountered in training such a large model. We place significant importance on the broader impacts of the release GPT-NeoX-20B and other such LLMs,

[1] This model appears to be under-trained and underperforms GPT-J-6B despite being twice the size.

and have prepared a separate manuscript for dissecting these issues in greater detail.

We are making the model weights at evenly spaced 1000 step intervals throughout the whole of training available, as well as open-sourcing the training and evaluation code, at [URL will go here after peer review]. We hope that by making a wide range of checkpoints throughout training freely available, we will facilitate research on the training dynamics of LLMs, as well as the aforementioned areas of AI safety and interpretability.

2 Model Design and Implementation

GPT-NeoX-20B is an autoregressive transformer decoder model whose architecture largely follows that of GPT-3 (Brown et al., 2020), with a few notable deviations described below. Our model has 20 billion parameters, of which 19.9 billion are "non-embedding" parameters that Kaplan et al. (2020) identify as the proper number to use for scaling laws analysis. Our model has 44 layers, a hidden dimension size of 6144, and 64 heads.

2.1 Model Architecture

Although our architecture is largely similar to GPT-3, there are some notable differences. In this section we give a high-level overview of those differences, but ask the reader to refer to (Brown et al., 2020) for full details of the model architecture. Our model architecture is almost identical to that of GPT-J (Wang and Komatsuzaki, 2021)[2], however we choose to use GPT-3 as the point of reference because there is no canonical published reference on the design of GPT-J.

2.1.1 Rotary Positional Embeddings

We use rotary embeddings (Su et al., 2021) instead of the learned positional embeddings used in GPT models (Radford et al., 2018), based on our positive prior experiences using it in training LLMs. Rotary embeddings are a form of static relative positional embeddings. In brief, they twist the embedding space such that the attention of a token at position m to token at position n is linearly dependent on $m-n$. More formally, they modify the standard multiheaded attention equations from

$$\text{softmax}\left(\frac{1}{\sqrt{d}}\sum_{n,m}\mathbf{x}_m^T\mathbf{W}_q^T\mathbf{W}_k\mathbf{x}_n\right),$$

[2]The unintentional difference is discussed in 2.2.2

where $\mathbf{x}_m, \mathbf{x}_n$ are (batched) embeddings of tokens at position m and n respectively and $\mathbf{W}_q^T, \mathbf{W}_k$ are the query and key weights respectively to

$$\text{softmax}\left(\frac{1}{\sqrt{d}}\sum_{n,m}\mathbf{x}_m^T\mathbf{W}_q^T R^d_{\Theta,(n-m)}\mathbf{W}_k\mathbf{x}_n\right),$$

where $R^d_{\Theta,x}$ is a $d\times d$ block diagonal matrix with the block of index i being a 2D rotation by $x\theta_i$ for hyperparameters $\Theta = \{\theta_i = 10000^{-2i/d} \mid i \in \{0,1,2,\ldots,(d-1)/2\}\}$. For a visual diagram of what rotary embeddings do, see Figure 1.

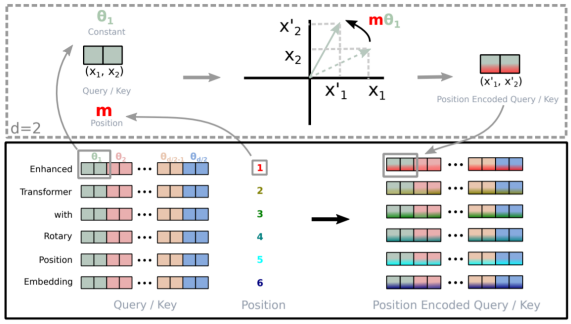


Figure 1: A pictorial representation of rotary embeddings, from Su et al. (2021).

While Su et al. (2021) apply rotary embeddings to every embedding vector, we follow Wang and Komatsuzaki (2021) and instead apply it only to the first 25% of embedding vector dimensions. Our initial experiments indicate that this strikes the best balance of performance and computational efficiency.[3]

2.1.2 Parallel Attention + FF Layers

We compute the Attention and Feed-Forward (FF) layers in parallel[4] and sum the results, rather than running them in series. This is primarily for efficiency purposes, as each residual addition with op-sharding requires one all-reduce in the forward pass and one in the backwards pass (Shoeybi et al., 2020). By computing the Attention and FFs in parallel, the results can be reduced locally before performing a single all-reduce. In Mesh Transformer JAX (Wang and Komatsuzaki, 2021), this led to a 15% throughput increase, while having comparable loss curves with running them in series during early training. See Section 2.2.2 for the full equation.

[3]Weights and Biases reports link provided after peer review.

[4][Code URL will go here after peer review] for implementation details.

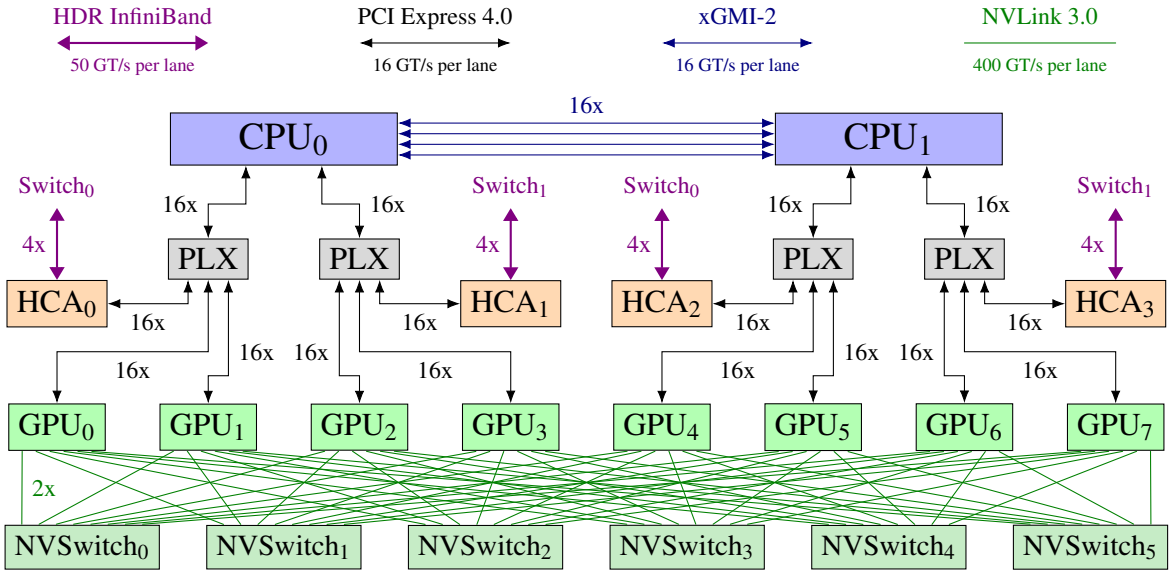


Figure 2: Architecture diagram of a single training node.

2.2 Initialization

For the Feed-Forward output layers before the residuals, we used the initialization scheme introduced in Wang (2021), $\frac{2}{L\sqrt{d}}$. This prevents activations from growing with increasing depth and width, with the factor of 2 compensating for the fact that the parallel and feed-forward layers are organized in parallel. For all other layers, we use the *small init* scheme from Nguyen and Salazar (2019), $\sqrt{\frac{2}{d+4d}}$

2.2.1 All Dense Layers

While GPT-3 uses alternating dense and sparse layers using the technique introduced in Child et al. (2019), we instead opt to exclusively use dense layers to reduce implementation complexity.

2.2.2 Double Layer Normalization

Due to an oversight, we unintentionally apply an extra Layer Norm in GPT-NeoX-20B. Instead of computing

$$x + \text{Attn}(\text{LN}(x)) + \text{FF}(\text{LN}(x))$$

as intended, our codebase computes

$$x + \text{Attn}(\text{LN}(\text{LN}(x))) + \text{FF}(\text{LN}(\text{LN}(x))).$$

Unfortunately, this was only noticed after we were much too far into training to restart. Subsequent experiments at small scales indicated that the extra layer norm makes no difference in performance, but we nevertheless wish to highlight this in the interest of transparency. We do not anticipate deliberately using this doubled Layer Norm in the future.

2.3 Software Libraries

Our model is trained using a codebase that builds on Megatron (Shoeybi et al., 2020) and DeepSpeed (Rasley et al., 2020) to facilitate efficient and straightforward training of large language models with tens of billions of parameters. We use the official PyTorch v1.10.0 release binary package compiled with CUDA 11.1. This package is bundled with NCCL 2.10.3 for distributed communications.

2.4 Hardware

We trained GPT-NeoX-20B on twelve Supermicro AS-4124GO-NART servers, each with eight NVIDIA A100-SXM4-40GB GPUs and configured with two AMD EPYC 7532 CPUs. All GPUs can directly access the InfiniBand switched fabric through one of four ConnectX-6 HCAs for GPUDirect RDMA. Two NVIDIA MQM8700-HS2R switches—connected by 16 links—compose the spine of this InfiniBand network, with one link per node CPU socket connected to each switch. Figure 2 shows a simplified overview of a node as configured for training.

3 Training

Due to the intractability of performing a hyperparameter sweep for a 20 billion parameter model, we opted to use the values from Brown et al. (2020) to guide our choice of hyperparameters. As Brown et al. (2020) did not train a model at our exact scale, we interpolate between the learning rates of their 13B and 175B models to arrive at a learning rate of 0.97E−5. Based on the results of smaller

scale experiments, we select a weight decay of 0.01. To achieve a higher training throughput, we opt to use the same batch size as OpenAI's 175B model–approximately 3.15M tokens, or 1538 contexts of 2048 tokens each, and train for a total of $150,000$ steps, decaying the learning rate with a cosine schedule to 10% of its original value at the end of training.

We use the AdamW (Loshchilov and Hutter, 2019) optimizer, with beta values of 0.9 and 0.95 respectively, and an epsilon of $1.0\text{E}{-8}$. We extend AdamW with the *ZeRO* optimizer (Rajbhandari et al., 2020) to reduce memory consumption by distributing optimizer states across ranks. Since the weights and optimizer states of a model at this scale do not fit on a single GPU, we use the tensor parallelism scheme introduced in Shoeybi et al. (2020) in combination with pipeline parallelism (Harlap et al., 2018) to distribute the model across GPUs. To train GPT-NeoX-20B, we found that the most efficient way to distribute the model given our hardware setup was to set a tensor parallel size of 2, and a pipeline parallel size of 4. This allows for the most communication intensive processes, tensor and pipeline parallelism, to occur within a node, and data parallel communication to occur across node boundaries. In this fashion, we were able to achieve and maintain an efficiency of 117 teraFLOPS per GPU.

3.1 Training Stability

When training large language models, ensuring stability throughout training is essential. We observed significant spikes in loss on several occasions that our model recovered from instantaneously. Manual inspection of these events indicate they represent highly anomalous sequences tokens that slipped through the filtering process, such as the same symbol repeated many times.

3.2 Training Data

GPT-NeoX-20B was trained on the Pile (Gao et al., 2020), a massive curated dataset designed specifically for training large language models. It consists of data from 22 data sources, coarsely broken down into 5 categories:

- **Academic Writing**: Pubmed Abstracts and PubMed Central, arXiv, FreeLaw,[5] USPTO Backgrounds,[6] PhilPapers,[7] NIH Exporter[8]
- **Web-scrapes and Internet Resources**: CommonCrawl, OpenWebText2, StackExchange,[9] Wikipedia (English)
- **Prose**: BookCorpus2, Bibliotik, Project Gutenberg (PG-19; Rae et al., 2019)
- **Dialogue**: Youtube subtitles, Ubuntu IRC,[10] OpenSubtitles (Lison and Tiedemann, 2016), Hacker News,[11] EuroParl (Koehn, 2005)
- **Miscellaneous**: GitHub, the DeepMind Mathematics dataset (Saxton et al., 2019), Enron Emails (Klimt and Yang, 2004)

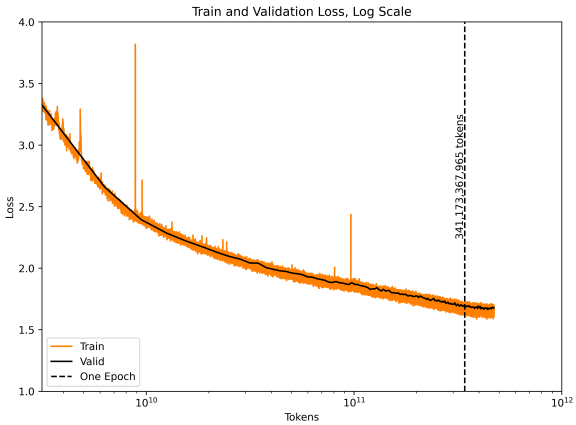


Figure 3: Training and validation loss for GPT-NeoX-20B. Following recent trends in language modeling, we aim to pretrain on only one epoch. As the validation loss continued to fall into the beginning of the second epoch, we decided to let it train for 1.5 epochs in hopes that the capacity gain of extra training would out-weigh the data redundancy.

In aggregate, the Pile consists of over 825GiB of raw text data. The diverse data sources reflects our desire for a general-purpose language model. Certain components are up-sampled to obtain a more balanced data distribution. In contrast, GPT-3's training data consists of web-scrapes, books datasets, and Wikipedia. When comparing results in this work to GPT-3, the training data is almost certainly the biggest known unknown factor. Full

[5] https://www.courtlistener.com/
[6] https://bulkdata.uspto.gov/
[7] https://philpapers.org/
[8] https://exporter.nih.gov/
[9] https://archive.org/details/stackexchange
[10] https://irclogs.ubuntu.com/
[11] https://news.ycombinator.com/

details of the Pile can be found in the technical report (Gao et al., 2020) and the associated datasheet (Biderman et al., 2022).

It is particularly notable that the Pile contains a scrape of StackExchange preprocessed into a Q/A form. There is a significant and growing body of work on the influence of the syntactic structure of finetuning data on downstream performance (Zhong et al., 2021; Tan et al., 2021; Sanh et al., 2021; Wei et al., 2021). While so far there has been no systematic work that focuses on *prompted pretraining*, recent work (Biderman and Raff, 2022) observed that the formulation of the StackExchange component of the Pile appears to heavily influences code generation. Biderman and Raff (2022) examine the generations produced by GPT-J-6B in response to prompts such as "write a Java program that accepts 10 integers and shows them in reversed order," finding not only a code solution but also a natural language discussion of that code, as one might see on Stack Exchange[12].

3.3 Tokenization

For GPT-NeoX-20B, we use a BPE-based tokenizer similar to that used in GPT-2, with the same total vocabulary size of 50257, with three major changes to the tokenizer. First, we train a new BPE tokenizer based on the Pile, taking advantage of its diverse text sources to construct a more general-purpose tokenizer. Second, in contrast to the GPT-2 tokenizer which treats tokenization at the start of a string as a non-space-delimited token, the GPT-NeoX-20B tokenizer applies consistent space delimitation regardless. This resolves an inconsistency regarding the presence of prefix spaces to a tokenization input.[13]. An example can be seen in Figure 4. Third, our tokenizer contains tokens for repeated space tokens (all positive integer amounts of repeated spaces up to and including 24). This allows the GPT-NeoX-20B tokenizer to tokenize text with large amounts of whitespace using fewer tokens; for instance, program source code or arXiv LaTeX source files.

3.3.1 Tokenizer Comparisons on Pretraining Corpora

Both tokenizers share 36938 out of 50257 tokens, a ∼73.5% overlap in tokens. In this section, we perform comparison between the GPT-NeoX-20B tokenizer to the GPT-2 tokenizer using the validation set of the Pile.

[12]In the cases the authors examine, the commentary on the solutions is factually incorrect. However we do not view this as problematic for the context of the current conversation.

[13]https://discuss.huggingface.co/t/bpe-tokenizers-and-spaces-before-words/475/2

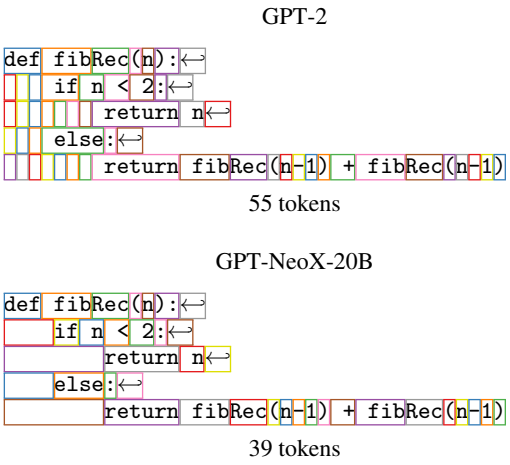


Figure 4: GPT-2 tokenization vs. GPT-NeoX-20B tokenization. GPT-NeoX-20B tokenization handles whitespace better, which is particularly useful for text such as source code. For more examples, see Appendix E.

In Table 9, we show the resulting number of tokens from tokenizing each component of the Piles validation set with both tokenizers, and the ratio of GPT-NeoX-20B tokens to GPT-2 tokens.

We observe that the GPT-NeoX-20B tokenizer represents all Pile components using fewer or very closely comparable numbers of tokens. The largest percentage improvement in token counts are in the EuroParl, GitHub, and PubMed Central components, with a more than 20% savings in the number of tokens needed to represent that component. We highlight that arXiv, GitHub, and StackExchange—subsets with large code components—can be represented with meaningfully fewer tokens with the GPT-NeoX-20B tokenizer compared to the GPT-2 tokenizer. Overall, the GPT-NeoX-20B tokenizer represents the Pile validation set with approximately 10% fewer tokens compared to the GPT-2 tokenizer.

Given that the GPT-NeoX-20B tokenizer is tweaked to better tokenize whitespace, we also perform a comparison between the two tokenizers excluding whitespace. We perform the same analysis as the above, but exclude all whitespace tokens from our computations, only counting the non-whitespace tokens. A token is considered a whitespace token if it consists only of whitespace characters. The results are shown in Table 10 in the Appendix. We observe that the GPT-NeoX-20B tokenizer still uses 5% fewer tokens to represent

the Pile validation set compared to the GPT-2 tokenizer. As expected, the token ratios for certain components such as GitHub and StackExchange become closer to even once the whitespace characters are excluded.

	GPT-2	GPT-NeoX-20B	$\frac{\text{GPT-NeoX-20B}}{\text{GPT-2}}$
Pile (val)	383,111,734	342,887,807	0.89501
C4	173,669,294	173,768,876	1.001
C4 excl. Space	168,932,391	171,003,008	1.012

Table 1: Number of tokens from tokenizing the AllenAI C4 (`en`) validation set. The GPT-NeoX-20B tokenizer uses approximately the same number of tokens to represent C4 as the GPT-2 tokenizer.

While we evaluated our tokenizer using the validation set for the Pile, the Pile components would still be considered in-domain for the tokenizer and may not provide the most informative comparison point. To perform an out-of-domain comparison, we perform the same analysis using the AllenAI replication of C4,[14], another popular pretraining corpus for large language models. As above, we use the validation set for our analysis. Our results are shown in Table 1. We find that the GPT-NeoX-20B tokenizer tokenizes the C4 validation set to approximately the same number of tokens as the GPT-2 tokenizer. When excluding all whitespace tokens, the GPT-NeoX-20B requires approximately 1% more tokens to represent the corpus compared to the GPT-2 tokenizer.

We include additional comparisons between the GPT-2 and GPT-NeoX-20B tokenizers in the Appendix, where we compare the longest tokens in each vocabulary (Table 11), worst-case tokenization examples (Table 12), and provide tokenization examples from each component of the Pile.

4 Performance Evaluations

To evaluate our model we use Gao et al. (2021b), an open source codebase for language model evaluation that supports a number of model APIs. We compare with the GPT-3 models on the OpenAI API[15] (Brown et al., 2020), the open source FairSeq dense models (Artetxe et al., 2021), and GPT-J-6B (Wang and Komatsuzaki, 2021). We do not compare against T5 (Raffel et al., 2020) as our evaluation methodology assumes that the models are autoregressive.

While it is common to display "scaling laws" curves of best fit, we opt to not do so as the small number of OpenAI API models give DaVinci an outsized influence on the slope of the curve. In many of the examples we study, including DaVinci in the scaling laws calculation moves the line of best fit so far as to entirely change the conclusions. Instead, we connect the points with lines directly. We categorize both GPT-J-6B and GPT-NeoX-20B under the umbrella of GPT-NeoX models, as both models are trained with the same architecture (except for the negligible differences described in 2.2.2) and were trained on the same dataset. However, we connect them using a dashed line to reflect the fact that these two models are not the same model trained at two different scales the way the FairSeq and OpenAI models are, having been trained using different codebases, different tokenizers, and for different numbers of tokens.

Where we were able to obtain the relevant information, we report two baselines: human-level performance and random performance. All plots contain error bars representing two standard errors, indicating the 95% confidence interval around each point. For some plots, the standard error is so small that the interval is not visible. Full results are shown in Appendix C.

4.1 Natural Language Tasks

We evaluate our model on a diverse collection of standard language model evaluation datasets. In general, we find that GPT-NeoX-20B performs comparably with FairSeq 13B on these datasets, and approximately on the linear interpolation between the performance of OpenAI's Curie and DaVinci models.

4.2 Advanced Knowledge-Based Tasks

In addition to evaluating on natural language processing benchmarks, we are also interested in the ability of models to answer factual questions requiring advanced knowledge. To do this, we use a dataset of multiple choice questions across diverse domains developed by Hendrycks et al. (2021a).

4.3 Mathematical Competency

Due to the fact that large language models tend to perform quite poorly on both arithmetic tasks and mathematical problems, we opted to include mathematical texts in various forms (arXiv, DM Mathe-

[14] https://github.com/allenai/allennlp/discussions/5056

[15] The numbers do not always agree with the numbers reported in Brown et al. (2020) due to slightly different prompt formatting.

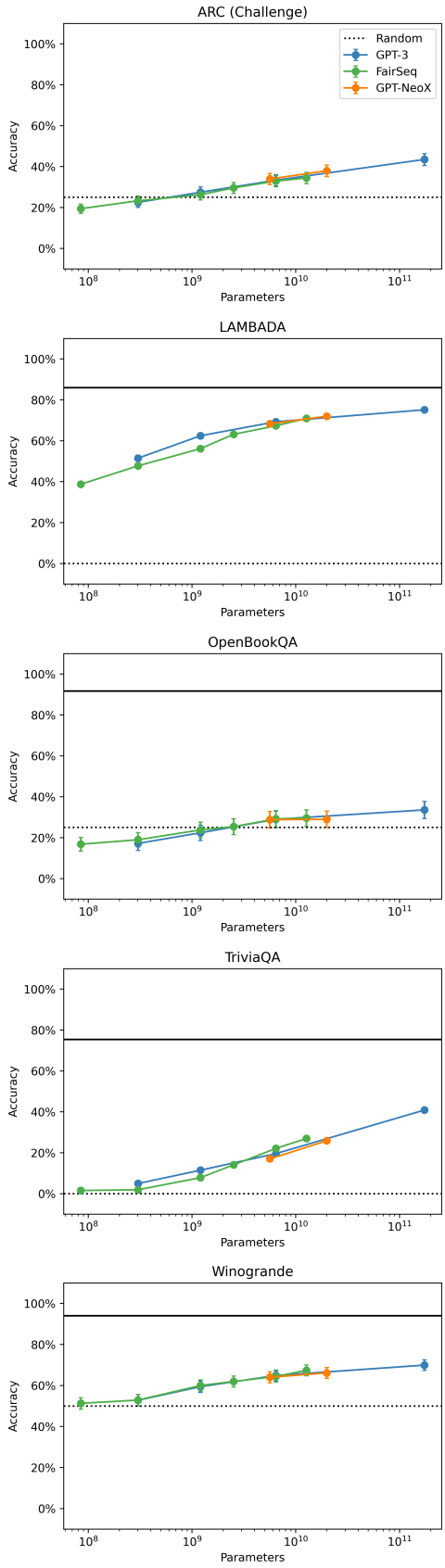


Figure 5: Zero-shot performance of GPT-NeoX-20B compared to GPT-J-6B and FairSeq and OpenAI models on a variety of language modeling benchmarks.

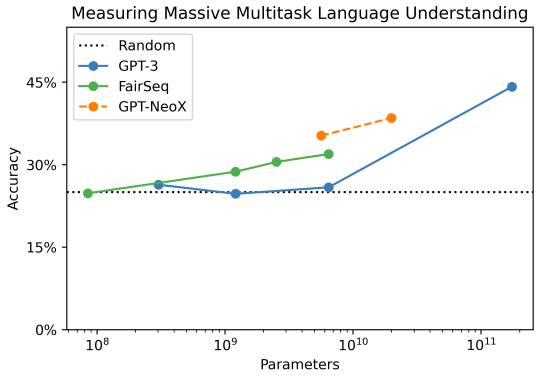


Figure 6: Five-shot performance of GPT-NeoX-20B compared to GPT-J-6B and FairSeq and OpenAI models on Hendrycks et al. (2021a).

matics, Math Stack Exchange and Math Overflow) as a significant portion of our training data in an attempt to improve performance in these areas.

We evaluate on the MATH test dataset (Hendrycks et al., 2021b) in Figure 7 Note that this is an evaluation metric that is typically fine-tuned on, but due to computational limitations we only evaluate models zero-shot here.

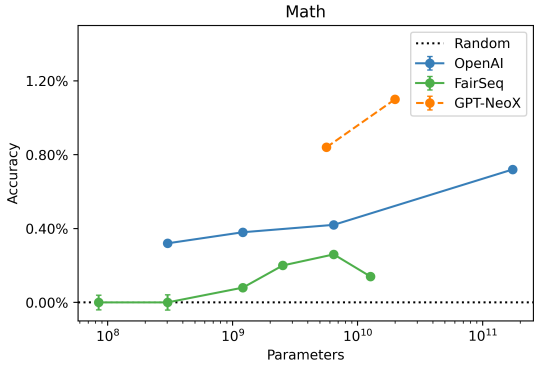


Figure 7: Zero-shot performance of GPT-NeoX-20B, GPT-J-6B, and GPT-Neo 2.7B compared to FairSeq and OpenAI models on the MATH dataset. Random performance on this task is 0%, and we were unable to find information on median human performance.

We also find that GPT-J-6B, also trained on the Pile, matches GPT-3 175B's performance despite being 30x smaller. While we leave finetuning GPT-J-6B and GPT-NeoX-20B to future work, we view this as a strong indicator that pretraining on the Pile is an effective way to improve performance on mathematics tasks.

5 Discussion

While the performance of the released 20B parameter model is impressive in many respects, outper-

forming GPT-J-6B, on most benchmarked datasets, the comparison to FairSeq's 13B parameter model, as well as the extrapolated performance based on OpenAI's models' performance, is a little more mixed. While GPT-NeoX-20B outperforms 13B on some tasks (e.g. ARC, LAMBADA, PIQA, PROST), it underperforms on others (e.g. HellaSwag, TriviaQA, Winogrande, WSC). On the other hand, performance in other areas such as scientific literature and mathematics exceeds that of compared models. In this section we try to dissect what the causes for these performance differences might be, and how we might be able to address them in future iterations.

Tokenizer and Dataset. We expect some performance regression on web- and book-based language modeling benchmarks due to our tokenizer design favoring scientific and programming documents [Section 3.3] over web documents, as well as the former being much more prevalent in the Pile when compared to GPT-3's training set. However, we lacked the resources to quantify the expected performance differences from these design choices, and leave this to future work.

Hyperparameter Tuning. Hyperparameter tuning is an expensive process, and is often infeasible to do at full scale for multi-billion parameter models. Due to the aforementioned limitations, we opted to choose hyperparameters based on a mixture of experiments at smaller scales and by interpolating parameters appropriate for our model size based on previously published work (Brown et al., 2020). However, several aspects of both our model architecture [Section 2.1] and training setup, including the data [Section 3.2] and the tokenizer [Section 3.3], diverge significantly from Brown et al. (2020). As such, it is almost certainly the case that the hyperparameters used for our model are no longer optimal, and potentially never were.

The effects of certain hyperparameter choices on language models are sadly understudied, to the extent that many aspects of training large transformer models are effectively a folk art, and only accessible to staff at a select number of private companies. For example, literature rigorously studying the effect of weight decay in language models is particularly scant (Zhou et al., 2021), (Zhang et al., 2018), and likewise, there is very little research investigating the effects of choices in designing a tokenizer—in particular the choice of vocabulary size—on downstream performance.

It is possible that our choice of 0.01 for weight decay was sub-optimal compared to GPT-3's choice of 0.1; however, we note that Fairseq's recently released LMs (Artetxe et al., 2021) also used a weight decay of 0.01. Another possible source of performance regression could have been our choice of batch size. For efficiency reasons, we opted to use a fairly large batch size of 3.2M tokens during training, the same as was used by the 175B GPT-3 model. The critical batch size (McCandlish et al., 2018) for the model may have been lower, possibly resulting in a negative impact on training. In addition, the optimal learning rate—which we selected by interpolating between the learning rates for models in (Brown et al., 2020)—can often be affected by the batch size, and failing to quantify this effect accurately could also have led to sub-optimal performance. During the training of the model, however, new research has been released which may significantly alleviate the costs of hyperparameter tuning for future training runs (Yang et al., 2021).

Dataset Deduplication. Finally, the lack of dataset deduplication could also have had an impact on downstream performance. Recent work has shown that deduplicating training data can have a large effect on perplexity (Lee et al., 2021). Although GPT-J-6B was trained on the same data and performed on par with an OpenAI model of a similar size, it is plausible that this effect could become more apparent with a larger model size.

6 Summary

We introduce GPT-NeoX-20B, a 20 billion parameter autoregressive Transformer language model trained on the Pile (Gao et al., 2020) dataset, and detail the main architectural differences between GPT-NeoX-20B and GPT-3—most notably the change in tokenizer, the addition of Rotary embeddings, the parallel computation of attention and feed-forward layers, and a different initialization scheme and hyperparameters. We run extensive evaluations of GPT-NeoX-20B on natural language and factual knowledge tasks, and compare it with other publicly available models, finding it performed particularly well on knowledge-based and mathematical tasks. Finally, we are open sourcing the training and evaluation code at [URL will go here after peer review], where readers can find a link to download the model weights across the whole training run.

References

Dario Amodei and Danny Hernandez. 2018. AI and compute. OpenAI Blog.

Stuart Armstrong and Sören Mindermann. 2018. Occam's razor is insufficient to infer the preferences of irrational agents. In *Advances in Neural Information Processing Systems*, volume 31, pages 5598–5609. Curran Associates, Inc.

Stuart Armstrong, Anders Sandberg, and Nick Bostrom. 2012. Thinking inside the box: Controlling and using an oracle AI. *Minds and Machines*, 22(4):299–324.

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 Diab, Zornitsa Kozareva, and Ves Stoyanov. 2021. Efficient large scale language modeling with mixtures of experts. *Computing Research Repository*, arXiv:2112.10684. Version 1.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. A general language assistant as a laboratory for alignment. *Computing Research Repository*, arXiv:2112.00861. Version 3.

Beth Barnes. 2021. Call for research on evaluating alignment (funding + advice available). AI Alignment Forum.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, New York, NY, USA. Association for Computing Machinery.

Tsvi Benson-Tilsen and Nate Soares. 2016. Formalizing convergent instrumental goals. In *AAAI Workshop: AI, Ethics, and Society*.

Asya Bergal and Nick Beckstead. 2021. Request for proposals for projects in AI alignment that work with deep learning systems. AI Alignment Forum.

Stella Biderman, Kieran Bicheno, and Leo Gao. 2022. Datasheet for the Pile. *Computing Research Repository*, arXiv:2201.07311. Version 1.

Stella Biderman and Edward Raff. 2022. Neural language models are effective plagiarists. *Computing Research Repository*, arXiv:2201.07406. Version 1.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow.

Nick Bostrom. 2014. *Superintelligence: Paths, Dangers, Strategies*, 1st edition. Oxford University Press, Inc., USA.

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.

Nick Cammarata, Shan Carter, Gabriel Goh, Chris Olah, Michael Petrov, Ludwig Schubert, Chelsea Voss, Ben Egan, and Swee Kiat Lim. 2020. Thread: Circuits. *Distill*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code. *Computing Research Repository*, arXiv:2107.03374. Version 2.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *Computing Research Repository*, arXiv:1904.10509. Version 1.

Paul Christiano. 2016. Prosaic AI Alignment. AI Alignment.

Paul Christiano, Ajeya Cotra, and Mark Xu. 2021. Eliciting latent knowledge: How to tell if your eyes deceive you.

Ajeya Cotra. 2020. Forecasting TAI with biological anchors.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, and Furu Wei. 2021. Knowledge neurons in pretrained transformers. *Computing Research Repository*, arXiv:2104.08696. Version 1.

Abram Demski. 2019. The parable of Predict-O-Matic. AI Alignment Forum.

Abram Demski and Scott Garrabrant. 2020. Embedded agency. *Computing Research Repository*, arXiv:1902.09469. Version 3.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *Computing Research Repository*, arXiv:1810.04805. Version 2.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2021. All NLP tasks are generation tasks: A general pretraining framework. *Computing Research Repository*, arXiv:2103.10360. Version 1.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A Mathematical Framework for Transformer Circuits. *transformer-circuits.pub*.

William Fedus, Barret Zoph, and Noam Shazeer. 2021. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Computing Research Repository*, arXiv:2101.03961. Version 1.

Joshua Fox and Carl Shulman. 2013. Superintelligence does not imply benevolence.

Deep Ganguli, Danny Hernandez, Liane Lovitt, Nova DasSarma, Tom Henighan, Andy Jones, Nicholas Joseph, Jackson Kernion, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Nelson Elhage, Sheer El Showk, Stanislav Fort, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, Neel Nanda, Kamal Ndousse, Catherine Olsson, Daniela Amodei, Dario Amodei, Tom Brown, Jared Kaplan, Sam McCandlish, Chris Olah, and Jack Clark. 2022. Predictability and surprise in large generative models. *Computing Research Repository*, arXiv:2202.07785. Version 1.

Leo Gao. 2021. Behavior cloning is miscalibrated. AI Alignment Forum.

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. 2020. The Pile: An 800GB dataset of diverse text for language modeling. *Computing Research Repository*, arXiv:2101.00027. Version 1.

Leo Gao, Kyle McDonell, Laria Reynolds, and Stella Biderman. 2021a. A preliminary exploration into factored cognition with language models. EleutherAI Blog.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021b. A framework for few-shot language model evaluation.

Scott Garrabrant, Tsvi Benson-Tilsen, Andrew Critch, Nate Soares, and Jessica Taylor. 2016. Logical induction. *Electron. Colloquium Comput. Complex.*, 23:154.

C. A. E. Goodhart. 1984. *Problems of Monetary Management: The UK Experience*, pages 91–121. Macmillan Education UK, London.

Katja Grace, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans. 2018. When will AI exceed human performance? evidence from AI experts. *Computing Research Repository*, arXiv:1705.08807. Version 3.

Aaron Harlap, Deepak Narayanan, Amar Phanishayee, Vivek Seshadri, Nikhil Devanur, Greg Ganger, and Phil Gibbons. 2018. PipeDream: Fast and efficient pipeline parallel DNN training. *Computing Research Repository*, arXiv:1806.03377. Version 1.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. *Computing Research Repository*, arXiv:2009.03300. Version 3.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the math dataset. *Computing Research Repository*, arXiv:2103.03874. Version 2.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. *Computing Research Repository*, arXiv:2201.07207. Version 1.

Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. 2021. Risks from learned optimization in advanced machine learning systems. *Computing Research Repository*, arXiv:1906.01820. Version 3.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Computing Research Repository*, arXiv:2001.08361. Version 1.

Holden Karnofsky. 2016. Some background on our views regarding advanced artificial intelligence. Open Philanthropy.

Bryan Klimt and Yiming Yang. 2004. The Enron corpus: A new dataset for email classification research. In *Proceedings of the 15th European Conference on Machine Learning*, ECML'04, page 217226, Berlin, Heidelberg. Springer-Verlag.

Jack Koch, Lauro Langosco, Jacob Pfau, James Le, and Lee Sharkey. 2021. Objective robustness in deep reinforcement learning. *Computing Research Repository*, arXiv:2105.14111. Version 2.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Vanessa Kosoy. 2016. IRL is hard. AI Alignment Forum.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *Computing Research Repository*, arXiv:1910.09700. Version 2.

Connor Leahy. 2021. Why Release a Large Language Model? EleutherAI Blog.

Connor Leahy and Stella Biderman. 2021. The hard problem of aligning AI to human values. In *The State of AI Ethics Report (Volume 4)*. The Montreal AI Ethics Institute.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2021. Deduplicating training data makes language models better. *Computing Research Repository*, arXiv:2107.06499. Version 1.

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

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. TruthfulQA: Measuring how models mimic human falsehoods. *Computing Research Repository*, arXiv:2109.07958. Version 1.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Computing Research Repository*, arXiv:1711.05101. Version 3.

David Manheim and Scott Garrabrant. 2019. Categorizing variants of goodhart's law. *Computing Research Repository*, arXiv:1803.04585. Version 4.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *Computing Research Repository*, arXiv:2005.00661. Version 1.

Sam McCandlish, Jared Kaplan, Dario Amodei, and OpenAI Dota Team. 2018. An empirical model of large-batch training. *Computing Research Repository*, arXiv:1812.06162. Version 1.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual knowledge in GPT. *Computing Research Repository*, arXiv:2202.05262v1. Version 1.

Toan Q. Nguyen and Julian Salazar. 2019. Transformers without tears: Improving the normalization of self-attention. *Computing Research Repository*, arXiv:1910.05895.

nostalgebraist. 2020. interpreting GPT: the logit lens. LessWrong.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. 2021. Show your work: Scratchpads for intermediate computation with language models. *Computing Research Repository*, arXiv:2112.00114. Version 1.

Stephen M. Omohundro. 2008. The basic ai drives. In *AGI*.

Pedro A. Ortega, Markus Kunesch, Grégoire Delétang, Tim Genewein, Jordi Grau-Moya, Joel Veness, Jonas Buchli, Jonas Degrave, Bilal Piot, Julien Perolat, Tom Everitt, Corentin Tallec, Emilio Parisotto, Tom Erez, Yutian Chen, Scott Reed, Marcus Hutter, Nando de Freitas, and Shane Legg. 2021. Shaking the foundations: delusions in sequence models for interaction and control. *Computing Research Repository*, arXiv:2110.10819. Version 1.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. Version 1.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. Technical report, OpenAI.

Alec Radford, Jeff 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. 2022. Scaling language models: Methods, analysis & insights from training Gopher. *Computing Research Repository*, arXiv:2112.11446. Version 2.

Jack W Rae, Anna Potapenko, Siddhant M Jayakumar, Chloe Hillier, and Timothy P Lillicrap. 2019. Compressive transformers for long-range sequence modelling. *Computing Research Repository*, arXiv:1911.05507. Version 1.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. ZeRO: Memory optimizations toward training trillion parameter models. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '20. IEEE Press.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. DeepSpeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506, New York, NY, USA. Association for Computing Machinery.

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 Févry, 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. *Computing Research Repository*, arXiv:2110.08207. Version 2.

David Saxton, Edward Grefenstette, Felix Hill, and Pushmeet Kohli. 2019. Analysing mathematical reasoning abilities of neural models. *Computing Research Repository*, arXiv:1904.01557. Version 1.

Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. 2020. Green AI. *Communications of the ACM*, 63(12):54–63.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2020. Megatron-lm: Training multi-billion parameter language models using model parallelism. *Computing Research Repository*, arXiv:1909.08053. Version 4.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, Elton Zhang, 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. *Computing Research Repository*, arXiv:2201.11990. Version 3.

Nate Soares. 2021. Visible thoughts project and bounty announcement. LessWrong.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. 2022. Learning to summarize from human feedback. *Computing Research Repository*, arXiv:2009.01325.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. 2021. Roformer: Enhanced transformer with rotary position embedding. *Computing Research Repository*, arXiv:2104.09864. Version 2.

Zhixing Tan, Xiangwen Zhang, Shuo Wang, and Yang Liu. 2021. MSP: Multi-stage prompting for making pre-trained language models better translators. *Computing Research Repository*, arXiv:2110.06609. Version 1.

Jie Tang. 2021. WuDao: Pretrain the world. Keynote address at the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases.

Jessica Taylor. 2016. Quantilizers: A safer alternative to maximizers for limited optimization. In *AAAI Workshop: AI, Ethics, and Society*.

Alexander Matt Turner, Logan Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli. 2021. Optimal policies tend to seek power. *Computing Research Repository*, arXiv:1912.01683.

Ben Wang. 2021. Mesh-Transformer-JAX: Model-Parallel Implementation of Transformer Language Model with JAX. `https://github.com/kingoflolz/mesh-transformer-jax`.

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

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. *Computing Research Repository*, arXiv:2109.01652. Version 5.

John Wentworth. 2020. Alignment by default. AI Alignment Forum.

Robert Wiblin. 2017. Positively shaping the development of artificial intelligence. 80k Hours Blog.

Ge Yang, Edward Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. 2021. Tuning large neural networks via zero-shot hyperparameter transfer. *Advances in Neural Information Processing Systems*, 34.

Eliezer Yudkowsky and Richard Ngo. 2021. Ngo and yudkowsky on alignment difficulty. LessWrong.

Eliezer Yudkowsky and Nate Soares. 2018. Functional decision theory: A new theory of instrumental rationality. *Computing Research Repository*, arXiv:1710.05060. Version 2.

Wei Zeng, Xiaozhe Ren, Teng Su, Hui Wang, Yi Liao, Zhiwei Wang, Xin Jiang, ZhenZhang Yang, Kaisheng Wang, Xiaoda Zhang, Chen Li, Ziyan Gong, Yifan Yao, Xinjing Huang, Jun Wang, Jianfeng Yu, Qi Guo, Yue Yu, Yan Zhang, Jin Wang, Hengtao Tao, Dasen Yan, Zexuan Yi, Fang Peng, Fangqing Jiang, Han Zhang, Lingfeng Deng, Yehong Zhang, Zhe Lin, Chao Zhang, Shaojie Zhang, Mingyue Guo, Shanzhi Gu, Gaojun Fan, Yaowei Wang, Xuefeng Jin, Qun Liu, and Yonghong Tian. 2021. Pangu-α: Large-scale autoregressive pretrained chinese language models with auto-parallel computation. *Computing Research Repository*, arXiv:2104.12369. Version 1.

Guodong Zhang, Chaoqi Wang, Bowen Xu, and Roger Grosse. 2018. Three mechanisms of weight decay regularization. *Computing Research Repository*, arXiv:2103.15345. Version 1.

Zhengyan Zhang, Yuxian Gu, Xu Han, Shengqi Chen, Chaojun Xiao, Zhenbo Sun, Yuan Yao, Fanchao Qi, Jian Guan, Pei Ke, Yanzheng Cai, Guoyang Zeng, Zhixing Tan, Zhiyuan Liu, Minlie Huang, Wentao Han, Yang Liu, Xiaoyan Zhu, and Maosong Sun. 2021. CPM-2: large-scale cost-effective pre-trained language models. *Computing Research Repository*, arXiv:2106.10715. Version 3.

Zhengyan Zhang, Xu Han, Hao Zhou, Pei Ke, Yuxian Gu, Deming Ye, Yujia Qin, YuSheng Su, Haozhe Ji, Jian Guan, Fanchao Qi, Xiaozhi Wang, Yanan Zheng, Guoyang Zeng, Huanqi Cao, Shengqi Chen, Daixuan Li, Zhenbo Sun, Zhiyuan Liu, Minlie Huang, Wentao Han, Jie Tang, Juanzi Li, Xiaoyan Zhu, and Maosong Sun. 2020. CPM: A large-scale generative chinese pre-trained language model. *Computing Research Repository*, arXiv:2012.00413. Version 1.

Ruiqi Zhong, Kristy Lee, Zheng Zhang, and Dan Klein. 2021. Adapting language models for zero-shot learning by meta-tuning on dataset and prompt collections. *Computing Research Repository*, arXiv:2104.04670. Version 5.

Yucong Zhou, Yunxiao Sun, and Zhao Zhong. 2021. FixNorm: Dissecting weight decay for training deep neural networks. *Computing Research Repository*, arXiv:2103.15345. Version 1.

A Full Configuration Details

In Table 2 we attach the full configuration details used to train GPT-NeoX-20B. The file is available in `.yaml` format usable in `gpt-neox` at [URL will go here after peer review], where we also provide documentation describing the role of each parameter.

Configuration Key	Value
attention-dropout	0
bias-gelu-fusion	True
checkpoint-activations	True
checkpoint-num-layers	1
data-impl	mmap
distributed-backend	nccl
eval-interval	1000
eval-iters	10
fp16.enabled	True
fp16.fp16	True
fp16.hysteresis	2
fp16.initial-scale-power	12
fp16.loss-scale	0
fp16.loss-scale-window	1000
fp16.min-loss-scale	1
gpt-j-residual	True
gradient-accumulation-steps	32
gradient-clipping	1.0
hidden-dropout	0
hidden-size	6144
init-method	small-init
log-interval	2
lr-decay-iters	150000
lr-decay-style	cosine
max-position-embeddings	2048
min-lr	9.7e-06
model-parallel-size	2
no-weight-tying	True
norm	layernorm
num-attention-heads	64
num-layers	44
optimizer.params.betas	[0.9, 0.95]
optimizer.params.eps	1e-08
optimizer.params.lr	9.7e-05
optimizer.type	Adam
output-layer-init-method	wang-init
output-layer-parallelism	column
partition-activations	False
pipe-parallel-size	4
pos-emb	rotary
rotary-pct	0.25
save-interval	500
scaled-upper-triang-masked-softmax-fusion	True
seq-length	2048
split	995,4,1
steps-per-print	2
synchronize-each-layer	True
tokenizer-type	HFTokenizer
train-iters	150000
train-micro-batch-size-per-gpu	4
vocab-file	20B-tokenizer.json
wall-clock-breakdown	False
warmup	0.01
weight-decay	0.01
zero-optimization.allgather-bucket-size	1260000000
zero-optimization.allgather-partitions	True
zero-optimization.contiguous-gradients	True
zero-optimization.cpu-offload	False
zero-optimization.overlap-comm	True
zero-optimization.reduce-bucket-size	1260000000
zero-optimization.reduce-scatter	True
zero-optimization.stage	1

Table 2: The full configuration details for GPT-NeoX-20B training

B Broader Impacts

We believe that Transformative Artificial Intelligence (TAI) (Karnofsky, 2016) is approaching (Cotra, 2020; Grace et al., 2018), and that these systems will cause catastrophic damage if they are misaligned with human values (Omohundro, 2008; Fox and Shulman, 2013). As such, we believe it is essential to prioritize and help facilitate technical research that ensures TAI's values will be aligned with ours.

AI Alignment generally refers to the problem of how to ensure increasingly powerful and autonomous AI systems perform the users' wishes faithfully and without unintended consequences. Alignment is especially critical as we approach human and superhuman levels of intelligence, as powerful optimization processes amplify small errors in goal specification into large misalignments (Goodhart, 1984; Fox and Shulman, 2013; Manheim and Garrabrant, 2019), and misalignments in this regime will result in runaway optimization processes that evade alteration or shutdown (Omohundro, 2008; Benson-Tilsen and Soares, 2016; Turner et al., 2021), posing a significant existential risk to humanity. Additionally, even if the goal is specified correctly, superhuman models may still develop deceptive subsystems that attempt to influence the real world to satisfy their objectives (Hubinger et al., 2021). While current systems are not yet at the level where the consequences of misalignment pose an existential threat, rapid progress in the field of AI has increased the concern that the alignment problem may be seriously tested in the not-too-distant future.

Much of the alignment literature focuses on the more theoretical aspects of alignment (Demski and Garrabrant, 2020; Yudkowsky and Soares, 2018; Taylor, 2016; Garrabrant et al., 2016; Armstrong and Mindermann, 2018; Hubinger et al., 2021), abstracting away the specifics of how intelligence will be implemented, due to uncertainty over the path to TAI. However, with the recent advances in capabilities, it may no longer be the case that the path to TAI is completely unpredictable. In particular, recent increases in the capabilities of large language models (LLMs) raise the possibility that the first generation of transformatively powerful AI systems may be based on similar principles and architectures as current large language models like GPT. This has motivated a number of research groups to work on "prosaic alignment" (Christiano,

2016; Askell et al., 2021; Ouyang et al., 2022), a field of study that considers the AI alignment problem in the case of TAI being built primarily with techniques already used in modern ML. We believe that due to the speed of AI progress, there is a significant chance that this assumption is true, and, therefore, that contributing and enabling contributions to prosaic alignment research will have a large impact.

The open-source release of this model is motivated by the hope that it will allow alignment researchers who would not otherwise have access to LLMs to use them. While there are negative risks due to the potential acceleration of capabilities research, which may place further time pressure on solving the alignment problem, we believe the benefits of this release outweigh the risks of accelerating capabilities research.

B.1 The Usefulness of Large Language Models in Alignment

LLMs represent a different paradigm than the AI systems generally studied by alignment researchers because they are not well-described as coherent agents or expected utility maximizers. Though trained to optimize a log-likelihood loss function, at a high level the goals a LLM pursues are varied and contradictory, depending on the way it is prompted. This introduces additional challenges, but may also enable new approaches to alignment.

GPT-NeoX-20B itself is not the system we need to align, but we hope it can serve as a publicly available platform for experiments whose results might generalize to crucial future work.

The following is a non-exhaustive list of potential approaches we consider promising for further investigation.

Mechanistic interpretability. Mechanistic interpretability research (Cammarata et al., 2020) hopes to gain an understanding into *how* models accomplish the tasks they do, in part in the hopes of detecting problematic or deceptive algorithms implemented by models before these failures manifest in the real world. Being able to interpret and inspect the detailed inner workings of trained models would be a powerful tool to ensure models are optimizing for the goals we intended (Hubinger et al., 2021; Koch et al., 2021). Reverse engineering transformer language models has already yielded insights about the inner functioning of LMs (Elhage et al., 2021; nostalgebraist, 2020; Meng et al., 2022; Dai et al., 2021).

Using a LLM as a reward model. Because they are trained to predict human writing, LLMs also appear to develop a useful representation of human values at the semantic level. Finding a way to utilise these representations could be a possible path toward solving the problem of reward robustness in RL and other algorithms which require a proxy of human judgment (Stiennon et al., 2022; Wentworth, 2020). Despite fundamental theoretical limitations on learning human values (Armstrong and Mindermann, 2018; Kosoy, 2016), value learning may still be robust enough to align weaker superhuman AIs. Future experiments could explore the extent to which LLM pretraining improves downstream reward model robustness and generalization.

Natural language transparency. Since LLM prompts are in a human-readable form, it can provide insight on the LLMs expected behavior. Prompt programming or finetuning can be used to leverage this fact and force a LLM to execute more transparent algorithms, such as splitting problems into steps or explicitly writing an "internal monologue" (Soares, 2021; Gao et al., 2021a; Nye et al., 2021). Reliability and trustworthiness can present significant challenges for these approaches.

However, this form of transparency also has its limits. In particular, models can often respond unpredictably to prompts, and internal monologues may become completely detached from the model's decision making process if translating between the model's ontology and the human ontology is more complex than simply modeling human monologues (Christiano et al., 2021).

Simulating agents at runtime. Although LLMs are not well-described as coherent agents, they can still be used to generate goal-directed processes. Given an appropriate prompt (such as a story of a character working to achieve a goal), LLMs can predict and thus simulate an agent (Huang et al., 2022). Simulated agents take representative actions according to the patterns present in the training data, similar to behavior cloning. One potential future research direction is testing whether they are less susceptible to failure modes that follow from expected utility maximization, such as Goodhart failures and power-seeking behavior. However, other failure modes can be introduced by the LM training procedure, such as "delusions" or "halluci-

nations" (Ortega et al., 2021; Gao, 2021; Maynez et al., 2020). Additionally, simulated agents may be uncompetitive with optimal agents like those produced by Reinforcement Learning. An important research direction is to explore how the beneficial properties of simulated agents can be maintained while making them competitive with RL based approaches.

Tool AI and automated alignment research. LMs can be used as relatively unagentic tools, such as OpenAI's Codex model (Chen et al., 2021) acting as a coding assistant. Because pretrained LLMs are not directly optimized for the factual accuracy of their predictions, it is possible they avoid some of the traditional problems with tool or oracle AI (Armstrong et al., 2012), such as the incentive to produce manipulative answers (Demski, 2019). Tool AI is not a long-term solution to the problem of alignment, but it could be used to assist alignment research or even automate large parts of it. For example, language models could be used to help brainstorm alignment ideas more quickly, act as a writing assistant, or directly generate alignment research papers for humans to review. This line of research also risks accelerating capabilities research, a concern we discuss more below.

B.2 Impact on Access to Large Language Models

The current status quo in research is that large language models are things people train and publish about, but do not actually release. To the best of our knowledge, the following is the 20 largest dense language models publicly disclosed[16]

B.3 Lack of Access

Despite the importance of prosaic alignment research, lack of access to large models presents a barrier to the types of research that can be explored by the wider research community. Having access to large models as close to the cutting edge as possible is essential for many research questions. Though some research can extrapolate from smaller models using scaling laws (Kaplan et al., 2020), some capabilities emerge only as models scale, and performance on tasks can increase discontinuously (Brown et al., 2020; Ganguli et al., 2022). Simply put, there are many properties of TAI you cannot study with models which can barely form coherent sentences.

Because training large models requires a significant engineering and capital investment, such models are often out of reach for small labs and independent researchers. As it stands, only large organizations have access to the latest generation of powerful language models (Brown et al., 2020; Rae et al., 2022; Fedus et al., 2021; Lieber et al., 2021; Tang, 2021). The number of researchers focused primarily on alignment working at these labs is much lower than those working on capabilities.

At current model capability levels, close-sourced models are primarily a limitation on the publics ability to probe and modify large models. Though some types of prosaic alignment research can be done using only inference through an API, many require direct access to network weights. For example, interpretability research relies heavily on the ability to inspect and modify weights and activations, and many approaches to control and robustness require additional training.

If only a small number of organizations and individuals have knowledge, access, and opportunity to work on pressing prosaic alignment problems, we are much more likely to fail. By openly training and releasing our large models, we hope to make it possible for more researchers to contribute to prosaic alignment. There are already a number of researchers using our previous models for alignment research (nostalgebraist, 2020; Meng et al., 2022; Lin et al., 2021), and there are multiple RFPs requesting more (Bergal and Beckstead, 2021; Barnes, 2021).

B.4 Impact on Capabilities Research

Alignment is, in a sense, "philosophy with a deadline" (Bostrom, 2014). Given the rapid progress of AI Capabilities, it seems likely that we have limited time to solve the alignment problem before existentially dangerous AI systems are developed. As it stands, alignment research lags far behind capabilities research and does not seem likely to catch up if the current rates of progress continue (Wiblin, 2017; Yudkowsky and Ngo, 2021; Amodei and Hernandez, 2018).

Given that it seems infeasible to halt all AI capabilities research, we reason that focusing on work that advances the field of alignment more than it advances AI capabilities should provide a net reduc-

[16] Mixture-of-Expert models have recently gained a significant amount of attention, but are not directly comparable to traditional dense transformers on a parameter-to-parameter or compute-to-compute basis and are excluded from this table.

tion to existential risk. We feel the risk of releasing GPT-NeoX-20B is acceptable, as the contribution of the model to capabilities research is likely to be limited, for two reasons. Firstly, the organizations pursuing capabilities research most aggressively are unlikely to benefit from our open-source release of this model because it significantly trails the state of the art. Second, we believe the single most impactful piece of knowledge that LLM research has had on advancing capabilities is the knowledge that scaling LMs was possible in the first place (Leahy, 2021; Leahy and Biderman, 2021), whereas the actual implementation is very fungible (as evidenced by the large number of parties who have succeeded in creating their own LLMs).

We ultimately believe that the benefits of releasing this model outweigh the risks, but this argument hinges crucially on the particular circumstances of this release. All actors considering releasing powerful AI models or advancing the frontier of capabilities should think carefully about what they release, in what way, and when.

B.5 Environmental Impact

A significant point of concern in some recent work is the energy usage and carbon emissions associated with training large language models (Strubell et al., 2019; Schwartz et al., 2020; Lacoste et al., 2019; Bender et al., 2021). In particular, Strubell et al. (2019) estimate that a then-recent paper by the authors released 626,155 lbs or 284.01 metric tons[17] of CO_2 (t_{CO_2}). As Strubell et al. (2019) has been widely cited and quoted in the media as representative of large-scale language models, we decided to explicitly and carefully track our energy usage and carbon emissions to see if this is truly a representative account of NLP emissions.

Throughout the development and training of our model, we tracked our energy usage and carbon emissions. We found that the process of developing and training GPT-NeoX-20B emitted almost exactly 10% of Strubell et al. (2019)'s estimate, coming in at a total of 69957 lbs or 31.73 metric tons of CO_2. This is roughly the equivalent of the yearly emissions of the average American or 35 round-trip flights between New York City and San Francisco. Our systems were based in Illinois, USA, and consumed energy sourced from the mix as follows

- 30.40% Coal ($0.95\,t_{CO_2}$/MWh)
- 31.30% Gas ($0.6078\,t_{CO_2}$/MWh)
- 1.30% Hydroelectric ($0\,t_{CO_2}$/MWh)
- 17.40% Nuclear ($0\,t_{CO_2}$/MWh)
- 0.30% Solar ($0\,t_{CO_2}$/MWh)
- 18.10% Wind ($0\,t_{CO_2}$/MWh)
- 1.30% Other Renewables ($0\,t_{CO_2}$/MWh)

This mixture produces an average of 0.47905 t_{CO_2}/MWh, and we consumed a total of 43.92 MWh of electricity over the course of 1830 hours of training. Scaling, testing, and evaluation were responsible for the equivalent of another 920 hours on our systems, for a total energy consumption 66.24 MWh and thus the production of just under 35 metric tons of CO_2.

It is noteworthy that Strubell et al. (2019) are estimating emissions from a *neural architecture search* paper, and is therefore not directly comparable to ours. The primary motivation for our comparison is that their number has attracted a lot of attention and is often taken to be respresentative of NLP research. In general, we advocate for more systematic and comprehensive reporting to improve transparency surrounding this important topic.

C Full Evaluation Results

Results for natural language understanding tasks are shown in Tables 3, 4, while results for Hendrycks tasks are found in Tables 5, 6, 7, 8.

All evaluations had version 0 in the Evaluation Harness. This information is reported in the output of the Evaluation Harness and should be used for ensuring reproducibility of these results, even as the task implementations themselves may change to fix bugs.

[17] We choose to present environmental impact figures in metric tons to align with standard reporting.

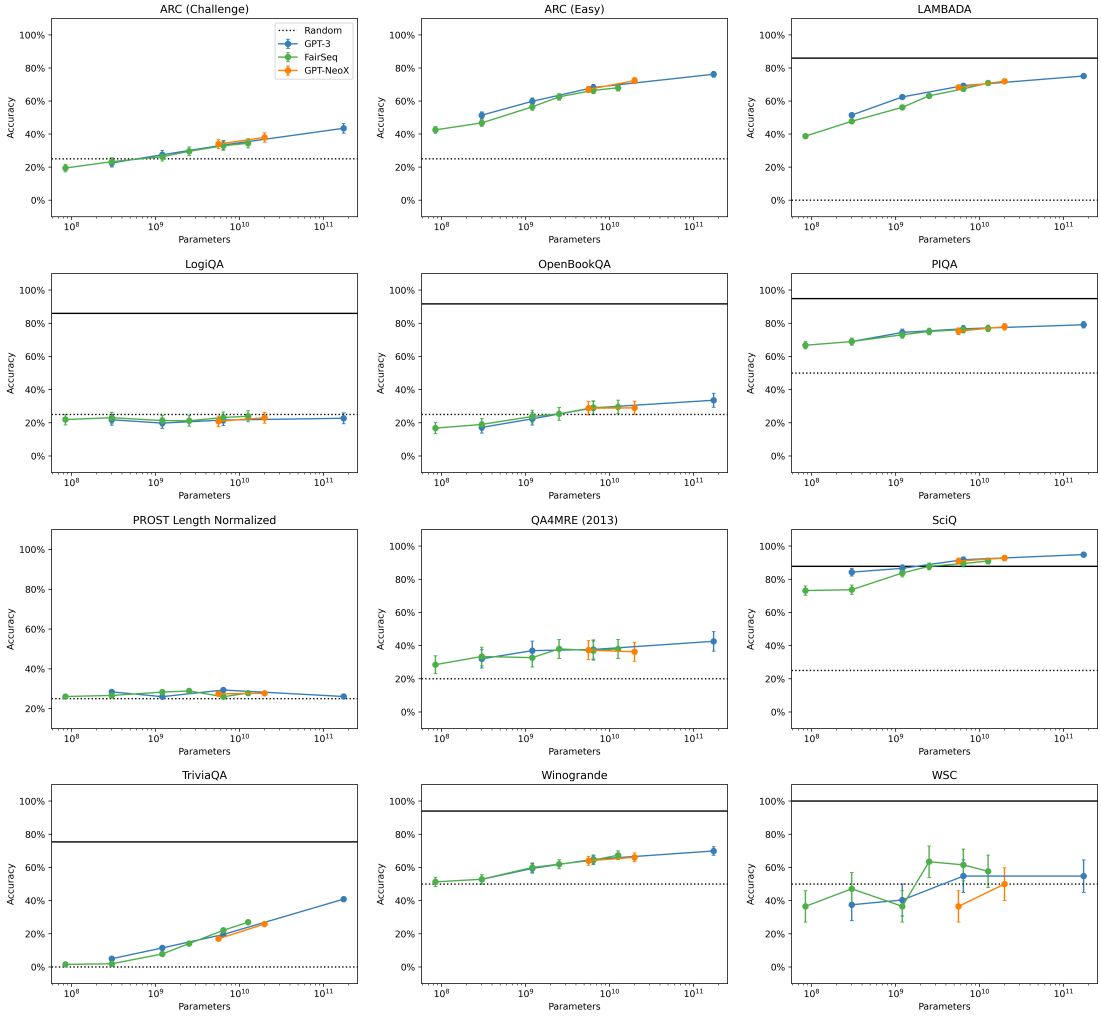


Figure 8: Zero-shot performance of GPT-NeoX-20B compared to GPT-J-6B and FairSeq and OpenAI models on a variety of language modeling benchmarks.

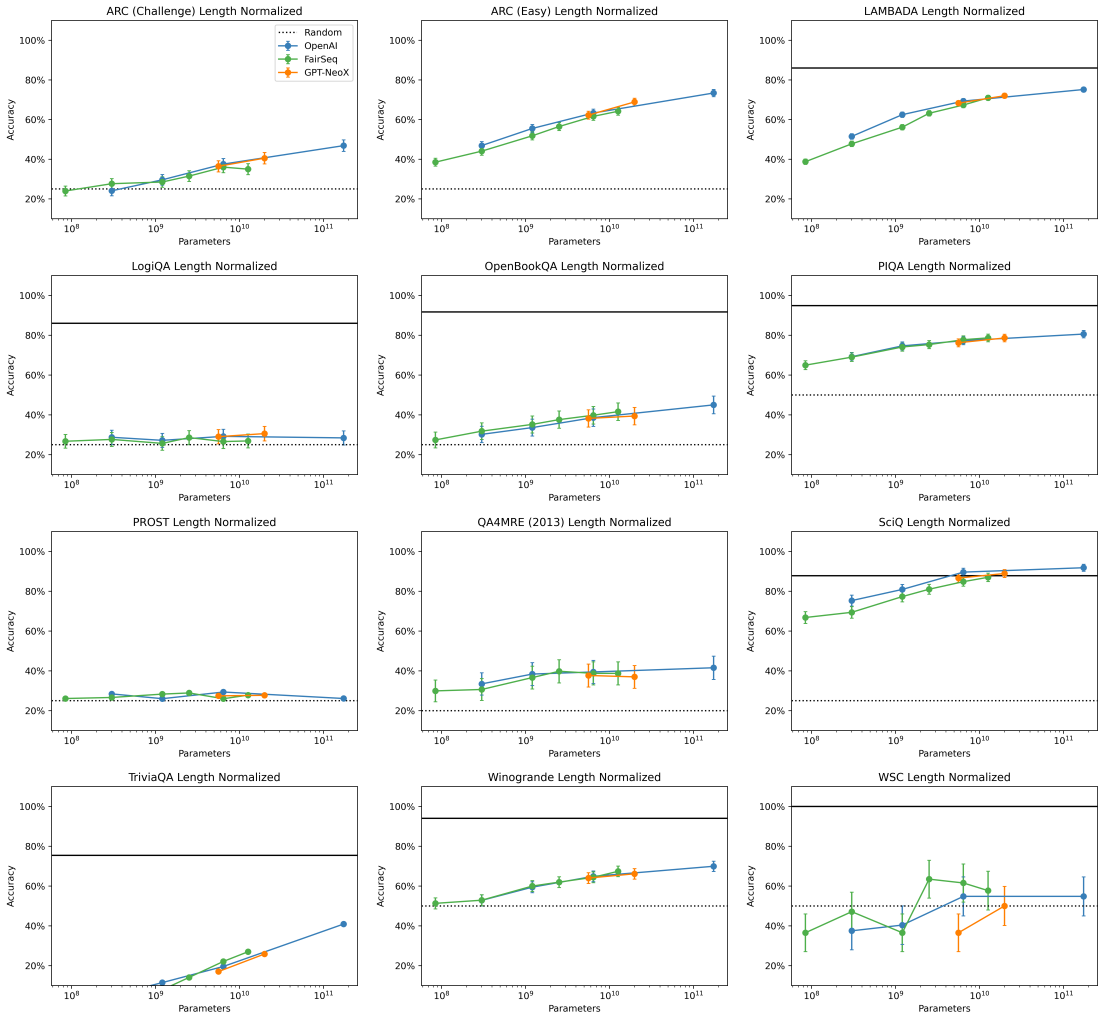


Figure 9: Length-normalized zero-shot performance of GPT-NeoX-20B compared to GPT-J-6B and FairSeq and OpenAI models on a variety of language modeling benchmarks.

Task	GPT-J 6B	GPT-NeoX 20B	GPT-3 Ada	GPT-3 Babbage	GPT-3 Curie	GPT-3 DaVinci
ANLI Round 1	0.324 ± 0.015	0.340 ± 0.015	0.334 ± 0.015	0.326 ± 0.015	0.325 ± 0.015	0.363 ± 0.015
ANLI Round 2	0.340 ± 0.015	0.343 ± 0.015	0.342 ± 0.015	0.308 ± 0.015	0.338 ± 0.015	0.375 ± 0.015
ANLI Round 3	0.355 ± 0.014	0.354 ± 0.014	0.354 ± 0.014	0.340 ± 0.014	0.353 ± 0.014	0.369 ± 0.014
LAMBADA	0.683 ± 0.006	0.720 ± 0.006	0.515 ± 0.007	0.625 ± 0.007	0.693 ± 0.006	0.752 ± 0.006
WSC	0.365 ± 0.047	0.500 ± 0.049	0.375 ± 0.048	0.404 ± 0.048	0.548 ± 0.049	0.548 ± 0.049
HellaSwag	0.518 ± 0.005	0.535 ± 0.005	0.359 ± 0.005	0.429 ± 0.005	0.505 ± 0.005	0.592 ± 0.005
Winogrande	0.640 ± 0.013	0.661 ± 0.013	0.528 ± 0.014	0.594 ± 0.014	0.649 ± 0.013	0.699 ± 0.013
SciQ	0.910 ± 0.009	0.928 ± 0.008	0.843 ± 0.012	0.866 ± 0.011	0.918 ± 0.009	0.949 ± 0.007
PIQA	0.752 ± 0.010	0.779 ± 0.010	0.690 ± 0.011	0.745 ± 0.010	0.767 ± 0.010	0.791 ± 0.009
TriviaQA	0.170 ± 0.004	0.259 ± 0.004	0.050 ± 0.002	0.115 ± 0.003	0.196 ± 0.004	0.409 ± 0.005
ARC (Easy)	0.670 ± 0.010	0.723 ± 0.009	0.514 ± 0.010	0.598 ± 0.010	0.682 ± 0.010	0.762 ± 0.009
ARC (Challenge)	0.340 ± 0.014	0.380 ± 0.014	0.225 ± 0.012	0.275 ± 0.013	0.334 ± 0.014	0.435 ± 0.014
OpenBookQA	0.288 ± 0.020	0.290 ± 0.020	0.172 ± 0.017	0.224 ± 0.019	0.290 ± 0.020	0.336 ± 0.021
HeadQA (English)	—	—	0.245 ± 0.008	0.278 ± 0.009	0.317 ± 0.009	0.356 ± 0.009
LogiQA	0.209 ± 0.016	0.230 ± 0.017	0.218 ± 0.016	0.198 ± 0.016	0.217 ± 0.016	0.227 ± 0.016
PROST	0.267 ± 0.003	0.296 ± 0.003	0.254 ± 0.003	0.270 ± 0.003	0.288 ± 0.003	0.267 ± 0.003
QA4MRE (2013)	0.373 ± 0.029	0.363 ± 0.029	0.320 ± 0.028	0.370 ± 0.029	0.377 ± 0.029	0.426 ± 0.029

Table 3: Results on Natural Language Understanding Tasks (GPT-NeoX and OpenAI Models)

	FairSeq					
Task	125M	355M	1.3B	2.7B	6.7B	13B
ANLI Round 1	0.316 ± 0.015	0.322 ± 0.015	0.331 ± 0.015	0.318 ± 0.015	0.338 ± 0.015	0.340 ± 0.015
ANLI Round 2	0.336 ± 0.015	0.312 ± 0.015	0.334 ± 0.015	0.339 ± 0.015	0.322 ± 0.015	0.330 ± 0.015
ANLI Round 3	0.330 ± 0.014	0.323 ± 0.014	0.333 ± 0.014	0.340 ± 0.014	0.333 ± 0.014	0.347 ± 0.014
LAMBADA	0.388 ± 0.007	0.478 ± 0.007	0.562 ± 0.007	0.632 ± 0.007	0.673 ± 0.007	0.709 ± 0.006
WSC	0.365 ± 0.047	0.471 ± 0.049	0.365 ± 0.047	0.635 ± 0.047	0.615 ± 0.048	0.577 ± 0.049
HellaSwag	0.309 ± 0.005	0.380 ± 0.005	0.448 ± 0.005	0.493 ± 0.005	0.525 ± 0.005	0.554 ± 0.005
Winogrande	0.513 ± 0.014	0.529 ± 0.014	0.600 ± 0.014	0.620 ± 0.014	0.644 ± 0.013	0.674 ± 0.013
SciQ	0.732 ± 0.014	0.737 ± 0.014	0.838 ± 0.012	0.878 ± 0.010	0.895 ± 0.010	0.910 ± 0.009
PIQA	0.668 ± 0.011	0.690 ± 0.011	0.731 ± 0.010	0.751 ± 0.010	0.762 ± 0.010	0.769 ± 0.010
TriviaQA	0.015 ± 0.001	0.019 ± 0.001	0.078 ± 0.003	0.141 ± 0.003	0.221 ± 0.004	0.270 ± 0.004
ARC (Easy)	0.426 ± 0.010	0.468 ± 0.010	0.565 ± 0.010	0.625 ± 0.010	0.665 ± 0.010	0.680 ± 0.010
ARC (Challenge)	0.195 ± 0.012	0.233 ± 0.012	0.263 ± 0.013	0.296 ± 0.013	0.329 ± 0.014	0.345 ± 0.014
OpenBookQA	0.168 ± 0.017	0.190 ± 0.018	0.238 ± 0.019	0.254 ± 0.019	0.292 ± 0.020	0.296 ± 0.020
HeadQA (English)	0.233 ± 0.008	0.233 ± 0.008	0.256 ± 0.008	0.264 ± 0.008	0.280 ± 0.009	0.280 ± 0.009
LogiQA	0.220 ± 0.016	0.230 ± 0.017	0.214 ± 0.016	0.212 ± 0.016	0.232 ± 0.017	0.240 ± 0.017
PROST	0.215 ± 0.003	0.257 ± 0.003	0.257 ± 0.003	0.230 ± 0.003	0.272 ± 0.003	0.252 ± 0.003
QA4MRE (2013)	0.285 ± 0.027	0.335 ± 0.028	0.327 ± 0.028	0.380 ± 0.029	0.370 ± 0.029	0.380 ± 0.029

Table 4: Results on Natural Language Understanding Tasks (FairSeq Models)

Task	GPT-J 6B	GPT-NeoX 20B	GPT-3 Ada	GPT-3 Babbage	GPT-3 Curie	GPT-3 DaVinci
Abstract Algebra	0.260 ± 0.044	0.230 ± 0.042	0.170 ± 0.038	0.220 ± 0.042	0.220 ± 0.042	0.220 ± 0.042
Anatomy	0.274 ± 0.039	0.319 ± 0.040	0.207 ± 0.035	0.289 ± 0.039	0.274 ± 0.039	0.348 ± 0.041
Astronomy	0.243 ± 0.035	0.329 ± 0.038	0.237 ± 0.035	0.211 ± 0.033	0.237 ± 0.035	0.382 ± 0.040
Business Ethics	0.290 ± 0.046	0.280 ± 0.045	0.360 ± 0.048	0.330 ± 0.047	0.300 ± 0.046	0.390 ± 0.049
Clinical Knowledge	0.272 ± 0.027	0.291 ± 0.028	0.223 ± 0.026	0.234 ± 0.026	0.253 ± 0.027	0.317 ± 0.029
College Biology	0.285 ± 0.038	0.271 ± 0.037	0.271 ± 0.037	0.299 ± 0.038	0.208 ± 0.034	0.347 ± 0.040
College Chemistry	0.240 ± 0.043	0.160 ± 0.037	0.270 ± 0.045	0.290 ± 0.046	0.210 ± 0.041	0.250 ± 0.044
College Computer Science	0.270 ± 0.045	0.250 ± 0.044	0.310 ± 0.046	0.270 ± 0.045	0.240 ± 0.043	0.260 ± 0.044
College Mathematics	0.260 ± 0.044	0.240 ± 0.043	0.220 ± 0.042	0.160 ± 0.037	0.200 ± 0.040	0.170 ± 0.038
College Medicine	0.197 ± 0.030	0.283 ± 0.034	0.237 ± 0.032	0.202 ± 0.031	0.225 ± 0.032	0.289 ± 0.035
College Physics	0.206 ± 0.040	0.284 ± 0.045	0.304 ± 0.046	0.324 ± 0.047	0.255 ± 0.043	0.235 ± 0.042
Computer Security	0.270 ± 0.045	0.290 ± 0.046	0.250 ± 0.044	0.240 ± 0.043	0.320 ± 0.047	0.350 ± 0.048
Conceptual Physics	0.255 ± 0.029	0.294 ± 0.030	0.264 ± 0.029	0.260 ± 0.029	0.268 ± 0.029	0.294 ± 0.030
Econometrics	0.237 ± 0.040	0.289 ± 0.043	0.289 ± 0.043	0.246 ± 0.040	0.246 ± 0.040	0.228 ± 0.039
Electrical Engineering	0.359 ± 0.040	0.303 ± 0.038	0.338 ± 0.039	0.276 ± 0.037	0.310 ± 0.039	0.414 ± 0.041
Elementary Mathematics	0.254 ± 0.022	0.283 ± 0.023	0.243 ± 0.022	0.272 ± 0.023	0.249 ± 0.022	0.312 ± 0.024
Formal Logic	0.341 ± 0.042	0.294 ± 0.041	0.262 ± 0.039	0.349 ± 0.043	0.270 ± 0.040	0.294 ± 0.041
Global Facts	0.250 ± 0.044	0.220 ± 0.042	0.240 ± 0.043	0.240 ± 0.043	0.300 ± 0.046	0.290 ± 0.046
High School Biology	0.252 ± 0.025	0.300 ± 0.026	0.235 ± 0.024	0.232 ± 0.024	0.271 ± 0.025	0.335 ± 0.027
High School Chemistry	0.202 ± 0.028	0.236 ± 0.030	0.246 ± 0.030	0.241 ± 0.030	0.197 ± 0.028	0.232 ± 0.030
High School Computer Science	0.250 ± 0.044	0.210 ± 0.041	0.190 ± 0.039	0.240 ± 0.043	0.220 ± 0.042	0.290 ± 0.046
High School European History	0.261 ± 0.034	0.255 ± 0.034	0.224 ± 0.033	0.285 ± 0.035	0.261 ± 0.034	0.303 ± 0.036
High School Geography	0.202 ± 0.029	0.227 ± 0.030	0.217 ± 0.029	0.207 ± 0.029	0.242 ± 0.031	0.348 ± 0.034
High School Government and Politics	0.228 ± 0.030	0.228 ± 0.030	0.212 ± 0.030	0.181 ± 0.028	0.212 ± 0.030	0.326 ± 0.034
High School Macroeconomics	0.285 ± 0.023	0.328 ± 0.024	0.272 ± 0.023	0.277 ± 0.023	0.277 ± 0.023	0.303 ± 0.023
High School Mathematics	0.219 ± 0.025	0.263 ± 0.027	0.196 ± 0.024	0.230 ± 0.026	0.167 ± 0.023	0.248 ± 0.026

Table 5: Results on Hendrycks Tasks, Part 1 (GPT-NeoX and OpenAI Models)

Task	GPT-J 6B	GPT-NeoX 20B	GPT-3 Ada	Babbage	Curie	DaVinci
High School Microeconomics	0.277 ± 0.029	0.294 ± 0.030	0.235 ± 0.028	0.265 ± 0.029	0.239 ± 0.028	0.307 ± 0.030
High School Physics	0.272 ± 0.036	0.298 ± 0.037	0.199 ± 0.033	0.298 ± 0.037	0.199 ± 0.033	0.219 ± 0.034
High School Physiology	0.273 ± 0.019	0.283 ± 0.019	0.209 ± 0.017	0.217 ± 0.018	0.246 ± 0.018	0.352 ± 0.020
High School Statistics	0.292 ± 0.031	0.319 ± 0.032	0.241 ± 0.029	0.278 ± 0.031	0.255 ± 0.030	0.278 ± 0.031
High School US History	0.289 ± 0.032	0.309 ± 0.032	0.255 ± 0.031	0.260 ± 0.031	0.240 ± 0.030	0.368 ± 0.034
High School World History	0.283 ± 0.029	0.295 ± 0.030	0.278 ± 0.029	0.262 ± 0.029	0.270 ± 0.029	0.321 ± 0.030
Human Aging	0.265 ± 0.030	0.224 ± 0.028	0.368 ± 0.032	0.336 ± 0.032	0.296 ± 0.031	0.327 ± 0.031
Human Sexuality	0.397 ± 0.043	0.405 ± 0.043	0.374 ± 0.042	0.427 ± 0.043	0.397 ± 0.043	0.481 ± 0.044
International Law	0.264 ± 0.040	0.298 ± 0.042	0.182 ± 0.035	0.207 ± 0.037	0.207 ± 0.037	0.331 ± 0.043
Jurisprudence	0.278 ± 0.043	0.250 ± 0.042	0.287 ± 0.044	0.278 ± 0.043	0.259 ± 0.042	0.370 ± 0.047
Logical Fallacies	0.294 ± 0.036	0.227 ± 0.033	0.239 ± 0.034	0.221 ± 0.033	0.245 ± 0.034	0.252 ± 0.034
Machine Learning	0.223 ± 0.040	0.268 ± 0.042	0.241 ± 0.041	0.286 ± 0.043	0.295 ± 0.043	0.232 ± 0.040
Management	0.233 ± 0.042	0.282 ± 0.045	0.184 ± 0.038	0.214 ± 0.041	0.320 ± 0.046	0.456 ± 0.049
Marketing	0.303 ± 0.030	0.321 ± 0.031	0.308 ± 0.030	0.282 ± 0.029	0.308 ± 0.030	0.491 ± 0.033
Medical Genetics	0.310 ± 0.046	0.340 ± 0.048	0.260 ± 0.044	0.300 ± 0.046	0.330 ± 0.047	0.430 ± 0.050
Miscellaneous	0.275 ± 0.016	0.299 ± 0.016	0.257 ± 0.016	0.269 ± 0.016	0.284 ± 0.016	0.450 ± 0.018
Moral Disputes	0.283 ± 0.024	0.289 ± 0.024	0.263 ± 0.024	0.263 ± 0.024	0.277 ± 0.024	0.301 ± 0.025
Moral Scenarios	0.237 ± 0.014	0.232 ± 0.014	0.238 ± 0.014	0.273 ± 0.015	0.238 ± 0.014	0.249 ± 0.014
Nutrition	0.346 ± 0.027	0.379 ± 0.028	0.301 ± 0.026	0.281 ± 0.026	0.291 ± 0.026	0.353 ± 0.027
Philosophy	0.260 ± 0.025	0.293 ± 0.026	0.215 ± 0.023	0.267 ± 0.025	0.244 ± 0.024	0.367 ± 0.027
Prehistory	0.244 ± 0.024	0.272 ± 0.025	0.244 ± 0.024	0.269 ± 0.025	0.284 ± 0.025	0.324 ± 0.026
Professional Accounting	0.262 ± 0.026	0.234 ± 0.025	0.202 ± 0.024	0.255 ± 0.026	0.238 ± 0.025	0.287 ± 0.027
Professional Law	0.241 ± 0.011	0.267 ± 0.011	0.261 ± 0.011	0.256 ± 0.011	0.259 ± 0.011	0.261 ± 0.011
Professional Medicine	0.276 ± 0.027	0.287 ± 0.027	0.221 ± 0.025	0.239 ± 0.026	0.265 ± 0.027	0.324 ± 0.028
Professional Psychology	0.284 ± 0.018	0.275 ± 0.018	0.245 ± 0.017	0.225 ± 0.017	0.257 ± 0.018	0.335 ± 0.019
Public Relations	0.282 ± 0.043	0.345 ± 0.046	0.255 ± 0.042	0.327 ± 0.045	0.364 ± 0.046	0.364 ± 0.046
Security Studies	0.363 ± 0.031	0.376 ± 0.031	0.367 ± 0.031	0.347 ± 0.030	0.384 ± 0.031	0.392 ± 0.031
Sociology	0.279 ± 0.032	0.284 ± 0.032	0.328 ± 0.033	0.303 ± 0.033	0.274 ± 0.032	0.368 ± 0.034
US Foreign Policy	0.340 ± 0.048	0.360 ± 0.048	0.330 ± 0.047	0.330 ± 0.047	0.380 ± 0.049	0.500 ± 0.050
Virology	0.355 ± 0.037	0.361 ± 0.037	0.307 ± 0.036	0.319 ± 0.036	0.337 ± 0.037	0.386 ± 0.038
World Religions	0.333 ± 0.036	0.386 ± 0.037	0.316 ± 0.036	0.310 ± 0.035	0.374 ± 0.037	0.398 ± 0.038

Table 6: Results on Hendrycks Tasks, Part 2 (GPT-NeoX and OpenAI Models)

	FairSeq					
Task	125M	355M	1.3B	2.7B	6.7B	13B
Abstract Algebra	0.260 ± 0.044	0.180 ± 0.039	0.230 ± 0.042	0.250 ± 0.044	0.240 ± 0.043	0.260 ± 0.044
Anatomy	0.178 ± 0.033	0.207 ± 0.035	0.185 ± 0.034	0.170 ± 0.032	0.259 ± 0.038	0.237 ± 0.037
Astronomy	0.270 ± 0.036	0.237 ± 0.035	0.243 ± 0.035	0.263 ± 0.036	0.296 ± 0.037	0.257 ± 0.036
Business Ethics	0.330 ± 0.047	0.410 ± 0.049	0.340 ± 0.048	0.350 ± 0.048	0.380 ± 0.049	0.340 ± 0.048
Clinical Knowledge	0.215 ± 0.025	0.264 ± 0.027	0.226 ± 0.026	0.249 ± 0.027	0.223 ± 0.026	0.264 ± 0.027
College Biology	0.285 ± 0.038	0.201 ± 0.034	0.243 ± 0.036	0.222 ± 0.035	0.271 ± 0.037	0.306 ± 0.039
College Chemistry	0.310 ± 0.046	0.290 ± 0.046	0.350 ± 0.048	0.300 ± 0.046	0.280 ± 0.045	0.240 ± 0.043
College Computer Science	0.200 ± 0.040	0.250 ± 0.044	0.260 ± 0.044	0.250 ± 0.044	0.300 ± 0.046	0.280 ± 0.045
College Mathematics	0.190 ± 0.039	0.170 ± 0.038	0.230 ± 0.042	0.200 ± 0.040	0.230 ± 0.042	0.250 ± 0.044
College Medicine	0.243 ± 0.033	0.237 ± 0.032	0.249 ± 0.033	0.254 ± 0.033	0.237 ± 0.032	0.260 ± 0.033
College Physics	0.216 ± 0.041	0.245 ± 0.043	0.216 ± 0.041	0.275 ± 0.044	0.343 ± 0.047	0.216 ± 0.041
Computer Security	0.240 ± 0.043	0.290 ± 0.046	0.300 ± 0.046	0.240 ± 0.043	0.230 ± 0.042	0.320 ± 0.047
Conceptual Physics	0.260 ± 0.029	0.255 ± 0.029	0.247 ± 0.028	0.243 ± 0.028	0.247 ± 0.028	0.204 ± 0.026
Econometrics	0.246 ± 0.040	0.272 ± 0.042	0.246 ± 0.040	0.281 ± 0.042	0.219 ± 0.039	0.263 ± 0.041
Electrical Engineering	0.283 ± 0.038	0.303 ± 0.038	0.234 ± 0.035	0.276 ± 0.037	0.310 ± 0.039	0.290 ± 0.038
Elementary Mathematics	0.246 ± 0.022	0.214 ± 0.021	0.233 ± 0.022	0.233 ± 0.022	0.246 ± 0.022	0.198 ± 0.021
Formal Logic	0.278 ± 0.040	0.302 ± 0.041	0.278 ± 0.040	0.310 ± 0.041	0.286 ± 0.040	0.333 ± 0.042
Global Facts	0.200 ± 0.040	0.210 ± 0.041	0.190 ± 0.039	0.150 ± 0.036	0.220 ± 0.042	0.160 ± 0.037
High School Biology	0.248 ± 0.025	0.255 ± 0.025	0.268 ± 0.025	0.226 ± 0.024	0.274 ± 0.025	0.235 ± 0.024
High School Chemistry	0.217 ± 0.029	0.207 ± 0.029	0.256 ± 0.031	0.281 ± 0.032	0.217 ± 0.029	0.266 ± 0.031
High School Computer Science	0.240 ± 0.043	0.230 ± 0.042	0.270 ± 0.045	0.240 ± 0.043	0.350 ± 0.048	0.280 ± 0.045
High School European History	0.230 ± 0.033	0.333 ± 0.037	0.279 ± 0.035	0.261 ± 0.034	0.273 ± 0.035	0.230 ± 0.033
High School Geography	0.263 ± 0.031	0.273 ± 0.032	0.222 ± 0.030	0.258 ± 0.031	0.207 ± 0.029	0.253 ± 0.031
High School Government and Politics	0.254 ± 0.031	0.290 ± 0.033	0.228 ± 0.030	0.233 ± 0.031	0.218 ± 0.030	0.187 ± 0.028
High School Macroeconomics	0.200 ± 0.020	0.272 ± 0.023	0.254 ± 0.022	0.269 ± 0.022	0.326 ± 0.024	0.256 ± 0.022
High School Mathematics	0.204 ± 0.025	0.189 ± 0.024	0.170 ± 0.023	0.226 ± 0.025	0.200 ± 0.024	0.193 ± 0.024

Table 7: Results on Hendrycks Tasks, Part 1 (FairSeq Models)

	FairSeq					
Task	125M	355M	1.3B	2.7B	6.7B	13B
High School Microeconomics	0.248 ± 0.028	0.256 ± 0.028	0.244 ± 0.028	0.248 ± 0.028	0.269 ± 0.029	0.227 ± 0.027
High School Physics	0.238 ± 0.035	0.219 ± 0.034	0.258 ± 0.036	0.245 ± 0.035	0.232 ± 0.034	0.166 ± 0.030
High School Physiology	0.235 ± 0.018	0.272 ± 0.019	0.266 ± 0.019	0.284 ± 0.019	0.250 ± 0.019	0.261 ± 0.019
High School Statistics	0.222 ± 0.028	0.241 ± 0.029	0.269 ± 0.030	0.250 ± 0.030	0.287 ± 0.031	0.241 ± 0.029
High School US History	0.240 ± 0.030	0.284 ± 0.032	0.299 ± 0.032	0.299 ± 0.032	0.314 ± 0.033	0.294 ± 0.032
High School World History	0.283 ± 0.029	0.232 ± 0.027	0.270 ± 0.029	0.245 ± 0.028	0.300 ± 0.030	0.316 ± 0.030
Human Aging	0.274 ± 0.030	0.309 ± 0.031	0.323 ± 0.031	0.291 ± 0.031	0.296 ± 0.031	0.274 ± 0.030
Human Sexuality	0.252 ± 0.038	0.366 ± 0.042	0.328 ± 0.041	0.359 ± 0.042	0.359 ± 0.042	0.351 ± 0.042
International Law	0.157 ± 0.033	0.223 ± 0.038	0.240 ± 0.039	0.281 ± 0.041	0.264 ± 0.040	0.231 ± 0.038
Jurisprudence	0.241 ± 0.041	0.269 ± 0.043	0.287 ± 0.044	0.241 ± 0.041	0.213 ± 0.040	0.278 ± 0.043
Logical Fallacies	0.196 ± 0.031	0.221 ± 0.033	0.233 ± 0.033	0.196 ± 0.031	0.245 ± 0.034	0.221 ± 0.033
Machine Learning	0.232 ± 0.040	0.295 ± 0.043	0.348 ± 0.045	0.232 ± 0.040	0.259 ± 0.042	0.241 ± 0.041
Management	0.223 ± 0.041	0.311 ± 0.046	0.214 ± 0.041	0.291 ± 0.045	0.340 ± 0.047	0.262 ± 0.044
Marketing	0.295 ± 0.030	0.231 ± 0.028	0.286 ± 0.030	0.303 ± 0.030	0.333 ± 0.031	0.329 ± 0.031
Medical Genetics	0.250 ± 0.044	0.310 ± 0.046	0.310 ± 0.046	0.280 ± 0.045	0.270 ± 0.045	0.300 ± 0.046
Miscellaneous	0.258 ± 0.016	0.301 ± 0.016	0.264 ± 0.016	0.249 ± 0.015	0.284 ± 0.016	0.268 ± 0.016
Moral Disputes	0.269 ± 0.024	0.246 ± 0.023	0.220 ± 0.022	0.260 ± 0.024	0.269 ± 0.024	0.272 ± 0.024
Moral Scenarios	0.255 ± 0.015	0.236 ± 0.014	0.273 ± 0.015	0.238 ± 0.014	0.241 ± 0.014	0.253 ± 0.015
Nutrition	0.252 ± 0.025	0.261 ± 0.025	0.297 ± 0.026	0.297 ± 0.026	0.330 ± 0.027	0.304 ± 0.026
Philosophy	0.199 ± 0.023	0.219 ± 0.023	0.228 ± 0.024	0.222 ± 0.024	0.238 ± 0.024	0.270 ± 0.025
Prehistory	0.290 ± 0.025	0.222 ± 0.023	0.253 ± 0.024	0.228 ± 0.023	0.296 ± 0.025	0.235 ± 0.024
Professional Accounting	0.262 ± 0.026	0.220 ± 0.025	0.209 ± 0.024	0.170 ± 0.022	0.238 ± 0.025	0.266 ± 0.026
Professional Law	0.261 ± 0.011	0.261 ± 0.011	0.256 ± 0.011	0.256 ± 0.011	0.259 ± 0.011	0.261 ± 0.011
Professional Medicine	0.239 ± 0.026	0.254 ± 0.026	0.254 ± 0.026	0.206 ± 0.025	0.221 ± 0.025	0.195 ± 0.024
Professional Psychology	0.245 ± 0.017	0.247 ± 0.017	0.242 ± 0.017	0.248 ± 0.017	0.278 ± 0.018	0.252 ± 0.018
Public Relations	0.236 ± 0.041	0.245 ± 0.041	0.264 ± 0.042	0.227 ± 0.040	0.291 ± 0.044	0.291 ± 0.044
Security Studies	0.322 ± 0.030	0.331 ± 0.030	0.331 ± 0.030	0.335 ± 0.030	0.408 ± 0.031	0.359 ± 0.031
Sociology	0.234 ± 0.030	0.234 ± 0.030	0.259 ± 0.031	0.229 ± 0.030	0.234 ± 0.030	0.323 ± 0.033
US Foreign Policy	0.250 ± 0.044	0.300 ± 0.046	0.300 ± 0.046	0.310 ± 0.046	0.370 ± 0.049	0.330 ± 0.047
Virology	0.289 ± 0.035	0.301 ± 0.036	0.319 ± 0.036	0.355 ± 0.037	0.295 ± 0.036	0.331 ± 0.037
World Religions	0.292 ± 0.035	0.263 ± 0.034	0.287 ± 0.035	0.292 ± 0.035	0.269 ± 0.034	0.339 ± 0.036

Table 8: Results on Hendrycks Tasks, Part 2 (FairSeq Models)

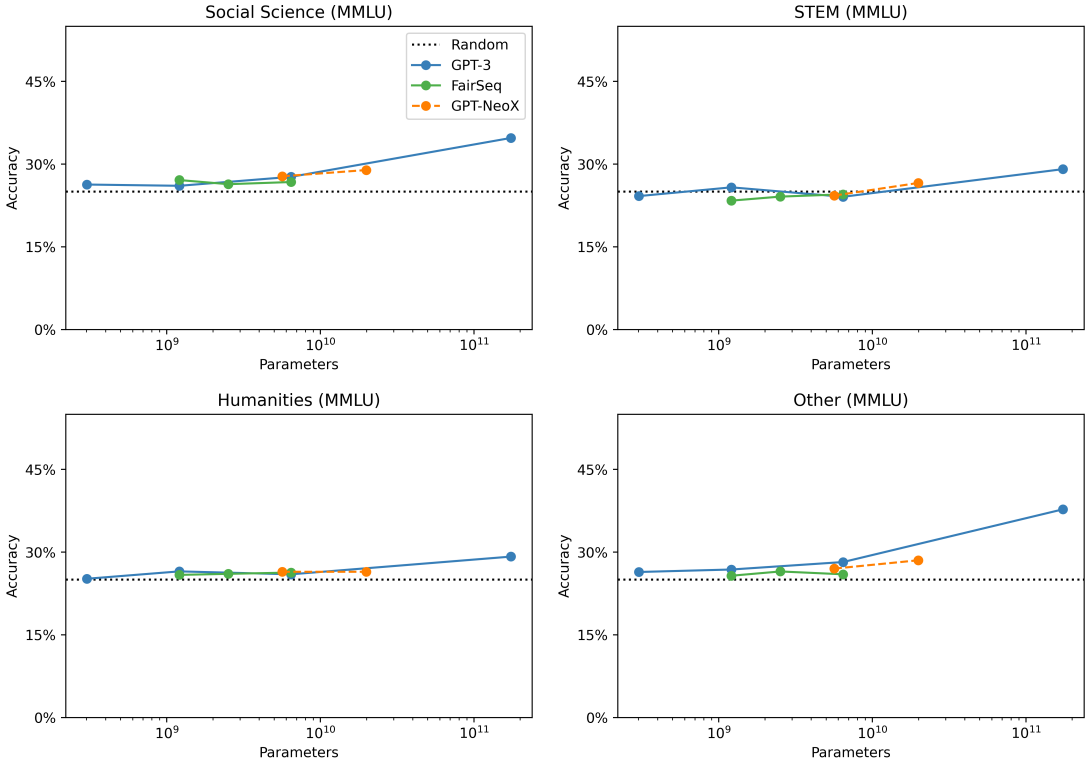


Figure 10: Zero-shot performance of GPT-NeoX-20B compared to GPT-J-6B and FairSeq and OpenAI models on Hendrycks et al. (2021a).

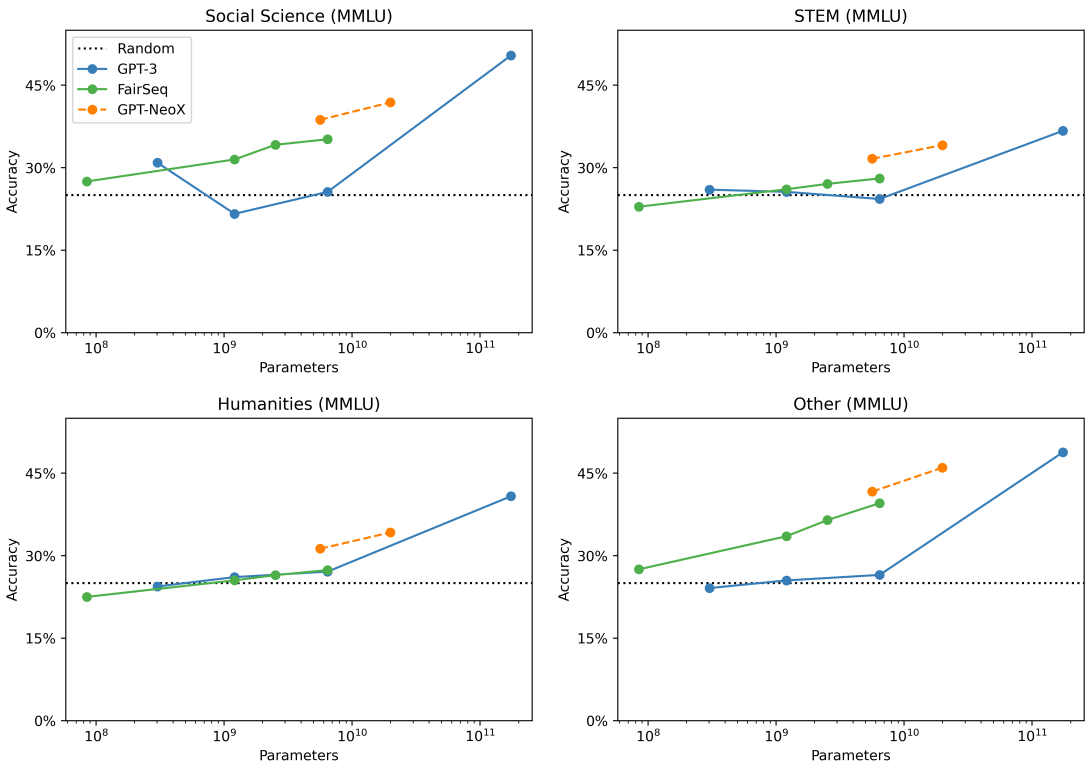


Figure 11: Five-shot performance of GPT-NeoX-20B compared to GPT-J-6B and FairSeq and OpenAI models on Hendrycks et al. (2021a). API limits we were unable to evaluate on the OpenAI API. Instead, we report numbers from Hendrycks et al. (2021a) with model sizes corrected.

D Additional Tokenizer Analysis

	GPT-2	GPT-NeoX-20B	$\frac{\text{GPT-NeoX-20B}}{\text{GPT-2}}$
arXiv	41,020,155	34,704,315	0.84603
BookCorpus2	2,336,388	2,365,633	1.01252
Books3	42,819,036	43,076,832	1.00602
DM Mathematics	7,699,527	7,413,775	0.96289
Enron Emails	480,500	433,867	0.90295
EuroParl	3,519,584	2,808,275	0.79790
FreeLaw	21,098,168	18,687,364	0.88573
GitHub	42,986,216	33,021,839	0.76820
Gutenberg (PG-19)	6,729,187	6,428,946	0.95538
HackerNews	2,578,933	2,551,720	0.98945
NIH ExPorter	776,688	739,558	0.95219
OpenSubtitles	5,431,529	5,446,485	1.00275
OpenWebText2	31,993,480	30,813,744	0.96313
PhilPapers	1,879,206	1,750,928	0.93174
Pile-CC	53,415,704	53,392,389	0.99956
PubMed Abstracts	8,708,180	8,215,529	0.94343
PubMed Central	56,874,247	43,534,166	0.76545
StackExchange	22,708,643	19,000,198	0.83669
USPTO Backgrounds	10,217,886	9,727,223	0.95198
Ubuntu IRC	3,341,287	2,771,066	0.82934
Wikipedia (en)	12,614,087	12,692,048	1.00618
YoutubeSubtitles	3,883,103	3,311,907	0.85290
Total	383,111,734	342,887,807	0.89501

Table 9: Number of tokens from tokenizing the Pile validation set. The GPT-NeoX-20B tokenizer uses fewer tokens to represent the Pile overall, with the biggest gains in whitespace heavy datasets such as arXiv, GitHub and StackExchange.

D.1 Tokenizer Comparisons

	GPT-2	GPT-NeoX-20B	$\frac{\text{GPT-NeoX-20B}}{\text{GPT-2}}$
arXiv	38,932,524	33,561,364	0.86204
BookCorpus2	2,233,367	2,262,609	1.01309
Books3	40,895,236	41,198,424	1.00741
DM Mathematics	7,214,874	6,929,066	0.96039
Enron Emails	374,978	373,498	0.99605
EuroParl	3,482,120	2,780,405	0.79848
FreeLaw	17,766,692	17,434,708	0.98131
GitHub	29,338,176	27,558,966	0.93936
Gutenberg (PG-19)	5,838,580	5,827,408	0.99809
HackerNews	2,312,116	2,299,848	0.99469
NIH ExPorter	776,619	739,543	0.95226
OpenSubtitles	5,428,118	5,445,721	1.00324
OpenWebText2	30,849,218	29,723,143	0.96350
PhilPapers	1,872,347	1,743,627	0.93125
Pile-CC	51,305,080	51,281,909	0.99955
PubMed Abstracts	8,676,790	8,185,417	0.94337
PubMed Central	44,508,570	40,722,151	0.91493
StackExchange	17,414,955	16,712,814	0.95968
USPTO Backgrounds	9,882,473	9,601,385	0.97156
Ubuntu IRC	3,220,797	2,659,225	0.82564
Wikipedia (en)	11,874,878	11,986,567	1.00941
YoutubeSubtitles	3,589,042	3,046,451	0.84882
Total	337,787,550	322,074,249	0.95348

Table 10: Number of tokens from tokenizing the Pile validation set, excluding whitespace tokens.

D.1.1 Longest Tokens

We show in Table 11 the 10 longest tokens in each tokenizer vocabulary. We exclude consideration of tokens that comprise only symbols or whitespace characters. We observe that for the GPT-2 tokenizer, many of the longest tokens appear to reflect artifacts in the tokenizer training data, likely with certain websites or web-scrapes being overrepresented in the training data. For the GPT-NeoX-20B tokenizer, we

observe that most of the longest tokens are scientific terms, likely arising from the PubMed components of the Pile.

GPT-2	GPT-NeoX-20B
rawdownloadcloneembedreportprint	immunohistochemistry
BuyableInstoreAndOnline	immunohistochemical
cloneembedreportprint	telecommunications
RandomRedditorWithNo	immunofluorescence
telecommunications	immunosuppressive
channelAvailability	BytePtrFromString
disproportionately	multidisciplinary
Telecommunications	histopathological
guiActiveUnfocused	neurodegenerative
ItemThumbnailImage	indistinguishable

Table 11: Ten longest tokens (excluding tokens comprising mainly symbols, numbers and spaces) in tokenizer vocabularies. “” indicates a word delimiter.

D.1.2 Worst Case Word Tokenization Comparison

GPT-2 Worst-case Tokenization

Word	GPT-2 Tokenization	GPT-NeoX-20B Tokenization
hematopoietic	(6) hematopoietic	(1) hematopoietic
adenocarcinoma	(6) adenocarcinoma	(1) adenocarcinoma
MERCHANTABILITY	(5) MERCHANTABILITY	(1) MERCHANTABILITY
CONSEQUENTIAL	(5) CONSEQUENTIAL	(1) CONSEQUENTIAL
oligonucleotides	(5) oligonucleotides	(1) oligonucleotides
cytoplasmic	(5) cytoplasmic	(1) cytoplasmic
corticosteroids	(4) corticosteroids	(1) corticosteroids
neurodegenerative	(4) neurodegenerative	(1) neurodegenerative
asymptotic	(4) asymptotic	(1) asymptotic
aneurysm	(4) aneurysm	(1) aneurysm

GPT-NeoX-20B Worst-case Tokenization

Word	GPT-2 Tokenization	GPT-NeoX-20B Tokenization
Schwarzenegger	(1) Schwarzenegger	(5) Schwarzenegger
Bolshevik	(1) Bolshevik	(4) Bolshevik
crowdfunding	(1) crowdfunding	(4) crowdfunding
misogyny	(1) misogyny	(4) misogyny
McAuliffe	(1) McAuliffe	(4) McAuliffe
unstoppable	(1) unstoppable	(4) unstoppable
Timberwolves	(1) Timberwolves	(4) Timberwolves
excruciating	(1) excruciating	(4) excruciating
Kaepernick	(1) Kaepernick	(4) Kaepernick
Valkyrie	(1) Valkyrie	(4) Valkyrie

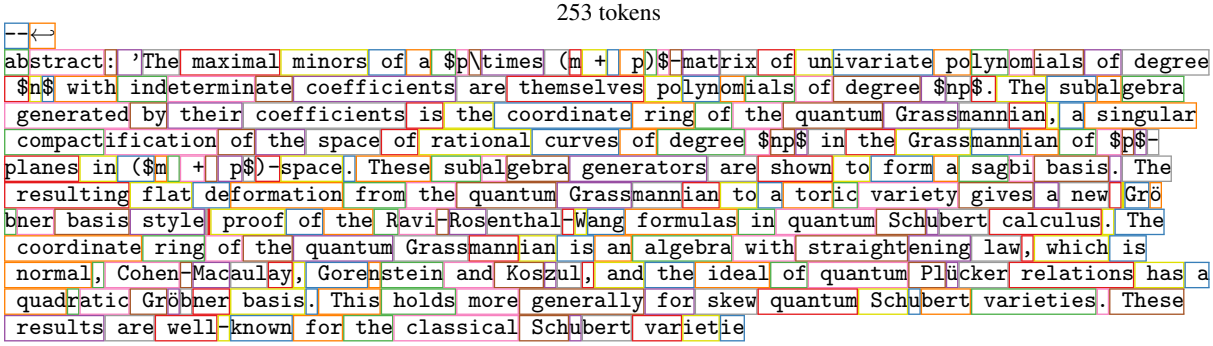
Table 12: Worst case word tokenization with respective tokenizers. We show cases where one tokenizer requires many more tokens to represent a word compared to the other tokenizer.

We consider the words for which there is the greatest discrepancy in the resulting token length between the two tokenizers, where one tokenizer needs many tokens to represent while the other tokenizer uses relatively few tokens. We define a word as a contiguous string delimited by whitespace or punctuation (as defined by `strings.punctuation` in Python). We perform this analysis at the component level. We only consider words that occur at least 10 times within the given component. We show in Table 12 a representative example from the Pile-CC corpus.

E Tokenization Examples

In Figures 12 - 17, we show examples of tokenized documents from the Pile, comparing the GPT-2 tokenizer to ours.

GPT-2 Tokenization
253 tokens



GPT-NeoX-20B Tokenization
229 tokens

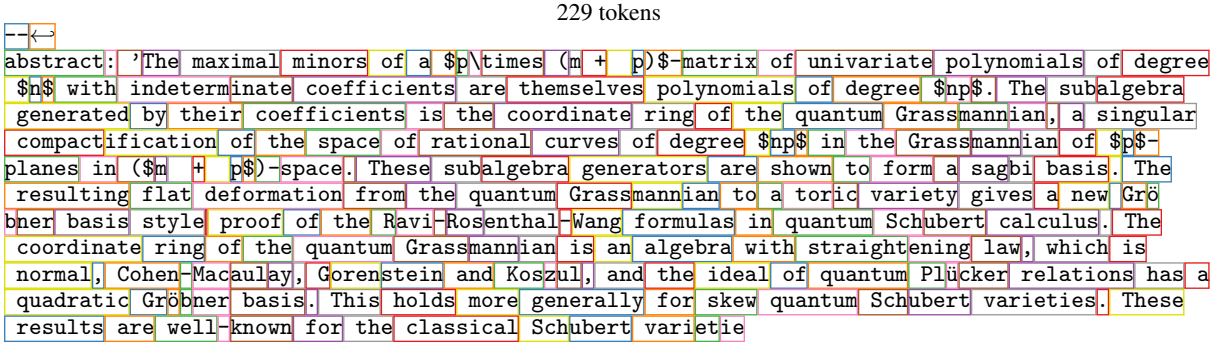
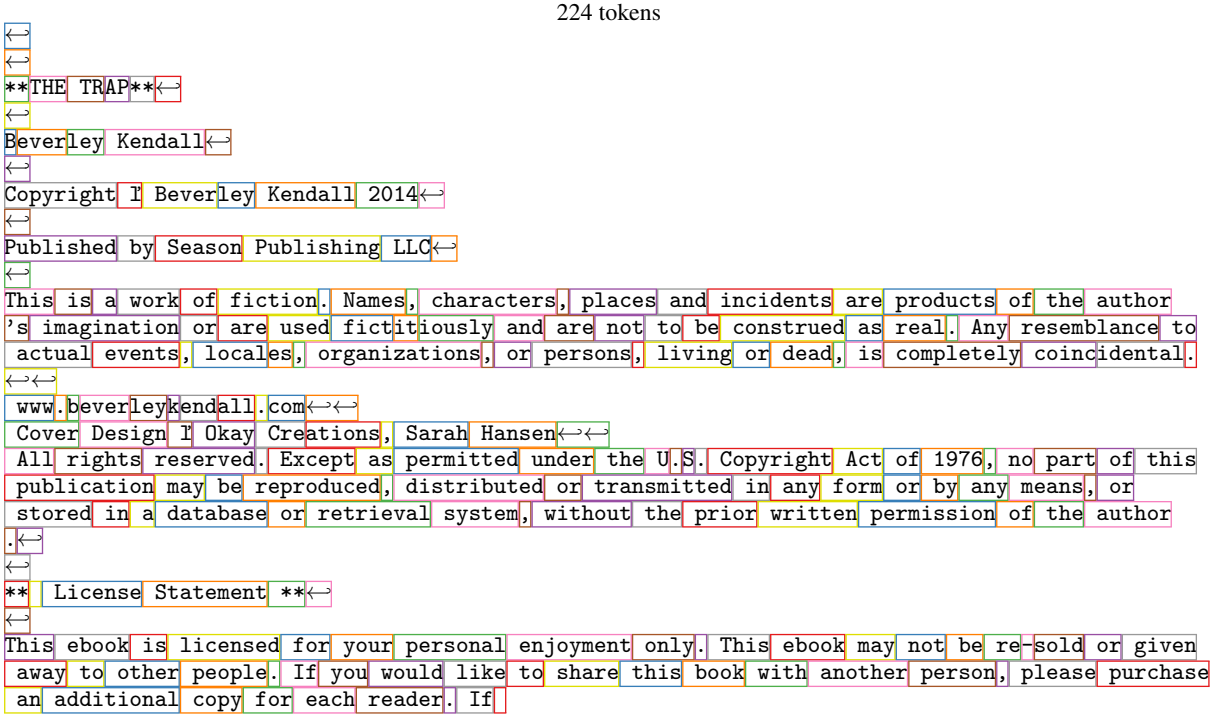


Figure 12: Pile (arXiv) Tokenization Example

GPT-2 Tokenization
224 tokens



GPT-NeoX-20B Tokenization
228 tokens

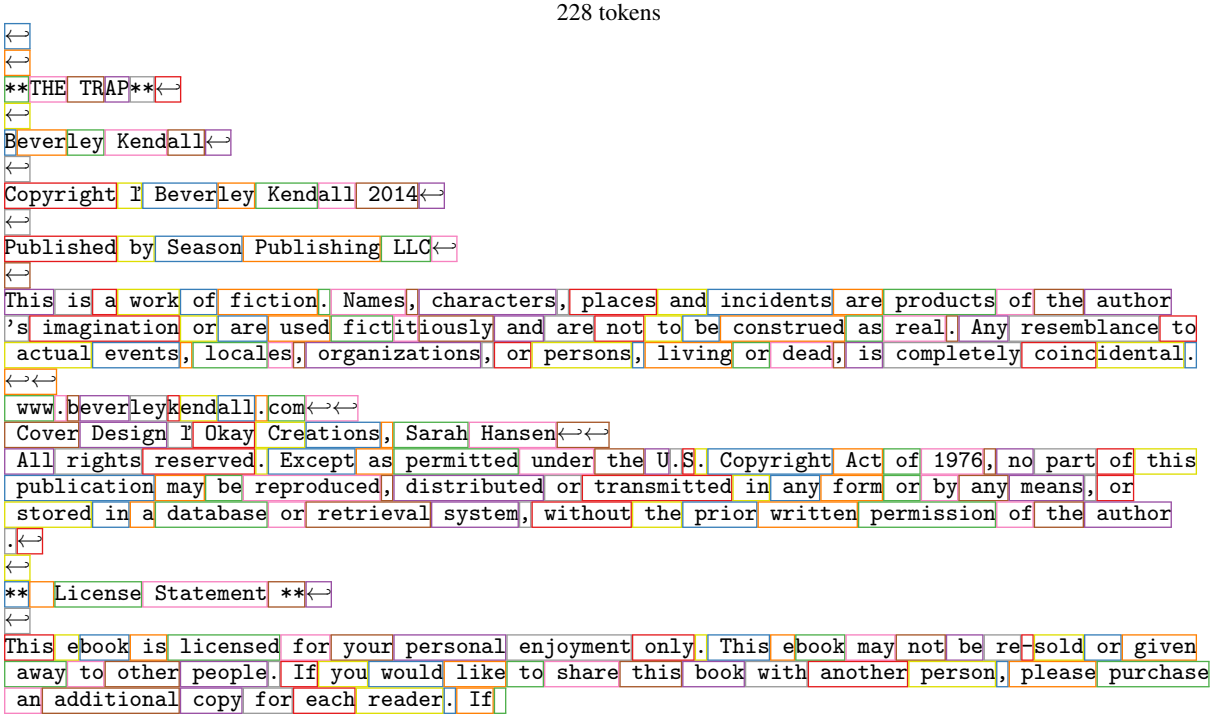


Figure 13: Pile (BookCorpus2) Tokenization Example

GPT-2 Tokenization
477 tokens

```
 o?↩
True↩
Suppose -3*t = 1 + 8. Let s(d) = d**3 + 6*d**2 + 2*d + 1. Let u be s(t). Suppose 10 = 5*z, 5*a +
 0*z = -z + u. Is 4 a factor of a?↩
True↩
Suppose 5*l = r - 35, -2*r + 5*l - 15 = -70. Is r a multiple of 4?↩
True↩
Suppose 2*l + 11 - 1 = 0. Does 15 divide (-2)/l - 118/(-5)?↩
False↩
Suppose 3*k - 3*f + 0*f - 72 = 0, -25 = -5*f. Is 9 a factor of 2/(-4) + k/2?↩
False↩
Suppose 6*w + 25 = w. Let t(c) = c + 9. Let u be t(w). Suppose -u*z = -3*z - 10. Is z a multiple
 of 5?↩
True↩
Let j = 81 + -139. Let i = j + 101. Is 11 a factor of i?↩
False↩
Let q(s) = s**3 + 4*s**2 - s + 2. Let u be q(-4). Let o(w) = w**2 + w - 6. Let t be o(u).
 Suppose -3*l - 39 = -3*d - 2*l, 0 = 3*d - 2*l - t. Does 9 divide d?↩
False↩
Suppose -2*b + 39 + 13 = 0. Is b a multiple of 14?↩
False↩
Let q = -7 + 12. Suppose 8*l = q*l + 81. Suppose 129 = 4*f - l. Is 13 a factor of f?↩
True↩
Suppose 0 = -4*n + j + 33, 4*n - n + 4*j = 20. Let c = 5 - n. Is 35*l - (-6)/c a multiple of 11?
↩
True↩
Let g(m) = m**2 - 2*m - 3. Let k be g(3). Let j be
```

GPT-NeoX-20B Tokenization
468 tokens

```
 o?↩
True↩
Suppose -3*t = 1 + 8. Let s(d) = d**3 + 6*d**2 + 2*d + 1. Let u be s(t). Suppose 10 = 5*z, 5*a +
 0*z = -z + u. Is 4 a factor of a?↩
True↩
Suppose 5*l = r - 35, -2*r + 5*l - 15 = -70. Is r a multiple of 4?↩
True↩
Suppose 2*l + 11 - 1 = 0. Does 15 divide (-2)/l - 118/(-5)?↩
False↩
Suppose 3*k - 3*f + 0*f - 72 = 0, -25 = -5*f. Is 9 a factor of 2/(-4) + k/2?↩
False↩
Suppose 6*w + 25 = w. Let t(c) = c + 9. Let u be t(w). Suppose -u*z = -3*z - 10. Is z a multiple
 of 5?↩
True↩
Let j = 81 + -139. Let i = j + 101. Is 11 a factor of i?↩
False↩
Let q(s) = s**3 + 4*s**2 - s + 2. Let u be q(-4). Let o(w) = w**2 + w - 6. Let t be o(u).
 Suppose -3*l - 39 = -3*d - 2*l, 0 = 3*d - 2*l - t. Does 9 divide d?↩
False↩
Suppose -2*b + 39 + 13 = 0. Is b a multiple of 14?↩
False↩
Let q = -7 + 12. Suppose 8*l = q*l + 81. Suppose 129 = 4*f - l. Is 13 a factor of f?↩
True↩
Suppose 0 = -4*n + j + 33, 4*n - n + 4*j = 20. Let c = 5 - n. Is 35*l - (-6)/c a multiple of 11?
↩
True↩
Let g(m) = m**2 - 2*m - 3. Let k be g(3). Let j be
```

Figure 14: Pile (DM Mathematics) Tokenization Example

GPT-2 Tokenization
430 tokens

```
<at-dialog title="vm.title" on-close="vm.onClose">↵
    <at-form state="vm.form" autocomplete="off" id="external_test_form">↵
        <at-input-group col="12" tab="20" state="vm.form.inputs" form-id="external_test"></at-
input-group>↵
        <at-action-group col="12" pos="right">↵
            <at-action-button↵
                variant="tertiary"↵
                ng-click="vm.onClose()"↵
            >↵
                ::vm.strings.get('CLOSE')↵
            </at-action-button>↵
            <at-action-button↵
                variant="primary"↵
                ng-click="vm.onSubmit()"↵
                ng-disabled="!vm.form.isValid || vm.form.disabled"↵
            >↵
                ::vm.strings.get('RUN')↵
            </at-action-button>↵
        </at-action-group>↵
    </at-form>↵
</at-dialog>↵
```

GPT-NeoX-20B Tokenization
257 tokens

```
<at-dialog title="vm.title" on-close="vm.onClose">↵
    <at-form state="vm.form" autocomplete="off" id="external_test_form">↵
        <at-input-group col="12" tab="20" state="vm.form.inputs" form-id="external_test"></at-
input-group>↵
        <at-action-group col="12" pos="right">↵
            <at-action-button↵
                variant="tertiary"↵
                ng-click="vm.onClose()"↵
            >↵
                ::vm.strings.get('CLOSE')↵
            </at-action-button>↵
            <at-action-button↵
                variant="primary"↵
                ng-click="vm.onSubmit()"↵
                ng-disabled="!vm.form.isValid || vm.form.disabled"↵
            >↵
                ::vm.strings.get('RUN')↵
            </at-action-button>↵
        </at-action-group>↵
    </at-form>↵
</at-dialog>↵
```

Figure 15: Pile (GitHub) Tokenization Example

GPT-2 Tokenization
178 tokens

Theresa May is expected to appoint an EU ambassador who believes in Brexit in the wake of the current Brussels representative's decision to quit after being cut adrift by Downing Street.

Sir Ivan Rogers on Tuesday announced his resignation as Britain's ambassador in Brussels after it was made clear Mrs May and her senior team had lost confidence in him over his pessimistic view of Brexit.

Government sources made clear that Sir Ivan had jumped before he was pushed and that Number 10 believed his negative view of Brexit meant that he could not lead the negotiations after the Prime Minister triggers Article 50.

In a 1,400-word resignation letter to his staff leaked on Tuesday night, Sir Ivan launched a thinly-veiled attack on the "muddled thinking" in Mrs May's Government.

GPT-NeoX-20B Tokenization
170 tokens

Theresa May is expected to appoint an EU ambassador who believes in Brexit in the wake of the current Brussels representative's decision to quit after being cut adrift by Downing Street.

Sir Ivan Rogers on Tuesday announced his resignation as Britain's ambassador in Brussels after it was made clear Mrs May and her senior team had lost confidence in him over his pessimistic view of Brexit.

Government sources made clear that Sir Ivan had jumped before he was pushed and that Number 10 believed his negative view of Brexit meant that he could not lead the negotiations after the Prime Minister triggers Article 50.

In a 1,400-word resignation letter to his staff leaked on Tuesday night, Sir Ivan launched a thinly-veiled attack on the "muddled thinking" in Mrs May's Government.

Figure 16: Pile (OpenWebText2) Tokenization Example

GPT-2 Tokenization
268 tokens

Carotid endarterectomy: operative risks, recurrent stenosis, and long-term stroke rates in a modern series.

To determine whether carotid endarterectomy (CEA) safely and effectively maintained a durable reduction in stroke complications over an extended period, we reviewed our data on 478 consecutive patients who underwent 544 CEA's since 1976. Follow-up was complete in 83% of patients (mean 44 months). There were 7 early deaths (1.3%), only 1 stroke related (0.2%). Peri operative stroke rates (overall 2.9%) varied according to operative indications: asymptomatic, 1 .4%; transient ischemic attacks (TIA)/amaurosis fugax (AF), 1.3%; nonhemispheric symptoms (NH), 4.9%; and prior stroke (CVA), 7.1%. Five and 10-year stroke-free rates were 96% and 92% in the asymptomatic group, 93% and 87% in the TIA/AF group, 92% and 92% in the NH group, and 80% and 73% in the CVA group. Late ipsilateral strokes occurred infrequently (8 patients, 1.7%). Late deaths were primarily cardiac related (51.3%). Stro

GPT-NeoX-20B Tokenization
250 tokens

Carotid endarterectomy: operative risks, recurrent stenosis, and long-term stroke rates in a modern series.

To determine whether carotid endarterectomy (CEA) safely and effectively maintained a durable reduction in stroke complications over an extended period, we reviewed our data on 478 consecutive patients who underwent 544 CEA's since 1976. Follow-up was complete in 83% of patients (mean 44 months). There were 7 early deaths (1.3%), only 1 stroke related (0.2%). Peri operative stroke rates (overall 2.9%) varied according to operative indications: asymptomatic, 1 .4%; transient ischemic attacks (TIA)/amaurosis fugax (AF), 1.3%; nonhemispheric symptoms (NH), 4.9%; and prior stroke (CVA), 7.1%. Five and 10-year stroke-free rates were 96% and 92% in the asymptomatic group, 93% and 87% in the TIA/AF group, 92% and 92% in the NH group, and 80% and 73% in the CVA group. Late ipsilateral strokes occurred infrequently (8 patients, 1.7%). Late deaths were primarily cardiac related (51.3%). Stro

Figure 17: Pile (PubMed Abstracts) Tokenization Example