EXHIBIT 177

Language Models are Unsupervised Multitask Learners

Alec Radford [*1] **Jeffrey Wu** [*1] **Rewon Child** [1] **David Luan** [1] **Dario Amodei** [**1] **Ilya Sutskever** [**1]

Abstract

Natural language processing tasks, such as question answering, machine translation, reading comprehension, and summarization, are typically approached with supervised learning on task-specific datasets. We demonstrate that language models begin to learn these tasks without any explicit supervision when trained on a new dataset of millions of webpages called WebText. When conditioned on a document plus questions, the answers generated by the language model reach 55 F1 on the CoQA dataset - matching or exceeding the performance of 3 out of 4 baseline systems without using the 127,000+ training examples. The capacity of the language model is essential to the success of zero-shot task transfer and increasing it improves performance in a log-linear fashion across tasks. Our largest model, GPT-2, is a 1.5B parameter Transformer that achieves state of the art results on 7 out of 8 tested language modeling datasets in a zero-shot setting but still underfits WebText. Samples from the model reflect these improvements and contain coherent paragraphs of text. These findings suggest a promising path towards building language processing systems which learn to perform tasks from their naturally occurring demonstrations.

1. Introduction

Machine learning systems now excel (in expectation) at tasks they are trained for by using a combination of large datasets, high-capacity models, and supervised learning (Krizhevsky et al., 2012) (Sutskever et al., 2014) (Amodei et al., 2016). Yet these systems are brittle and sensitive to slight changes in the data distribution (Recht et al., 2018) and task specification (Kirkpatrick et al., 2017). Current systems are better characterized as narrow experts rather than competent generalists. We would like to move towards more general systems which can perform many tasks – eventually without the need to manually create and label a training dataset for each one.

The dominant approach to creating ML systems is to collect a dataset of training examples demonstrating correct behavior for a desired task, train a system to imitate these behaviors, and then test its performance on independent and identically distributed (IID) held-out examples. This has served well to make progress on narrow experts. But the often erratic behavior of captioning models (Lake et al., 2017), reading comprehension systems (Jia & Liang, 2017), and image classifiers (Alcorn et al., 2018) on the diversity and variety of possible inputs highlights some of the shortcomings of this approach.

Our suspicion is that the prevalence of single task training on single domain datasets is a major contributor to the lack of generalization observed in current systems. Progress towards robust systems with current architectures is likely to require training and measuring performance on a wide range of domains and tasks. Recently, several benchmarks have been proposed such as GLUE (Wang et al., 2018) and decaNLP (McCann et al., 2018) to begin studying this.

Multitask learning (Caruana, 1997) is a promising framework for improving general performance. However, multitask training in NLP is still nascent. Recent work reports modest performance improvements (Yogatama et al., 2019) and the two most ambitious efforts to date have trained on a total of 10 and 17 `(dataset, objective)` pairs respectively (McCann et al., 2018) (Bowman et al., 2018). From a meta-learning perspective, each `(dataset, objective)` pair is a single training example sampled from the distribution of datasets and objectives. Current ML systems need hundreds to thousands of examples to induce functions which generalize well. This suggests that multitask training many need just as many effective training pairs to realize its promise with current approaches. It will be very difficult to continue to scale the creation of datasets and the design of objectives to the degree that may be required to brute force our way there with current techniques. This motivates exploring additional setups for performing multitask learning.

The current best performing systems on language tasks

[*,**]Equal contribution [1]OpenAI, San Francisco, California, United States. Correspondence to: Alec Radford <alec@openai.com>.

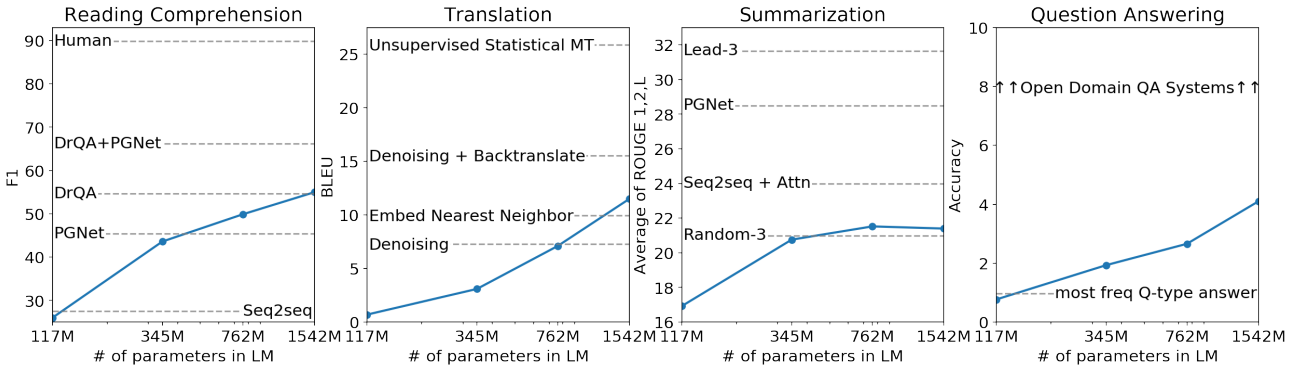


Figure 1. Zero-shot task performance of WebText LMs as a function of model size on many NLP tasks. Reading Comprehension results are on CoQA (Reddy et al., 2018), translation on WMT-14 Fr-En (Artetxe et al., 2017), summarization on CNN and Daily Mail (See et al., 2017), and Question Answering on Natural Questions (Kwiatkowski et al., 2019). Section 3 contains detailed descriptions of each result.

utilize a combination of pre-training and supervised fine-tuning. This approach has a long history with a trend towards more flexible forms of transfer. First, word vectors were learned and used as inputs to task-specific architectures (Mikolov et al., 2013) (Collobert et al., 2011), then the contextual representations of recurrent networks were transferred (Dai & Le, 2015) (Peters et al., 2018), and recent work suggests that task-specific architectures are no longer necessary and transferring many self-attention blocks is sufficient (Radford et al., 2018) (Devlin et al., 2018).

These methods still require supervised training in order to perform a task. When only minimal or no supervised data is available, another line of work has demonstrated the promise of language models to perform specific tasks, such as commonsense reasoning (Schwartz et al., 2017) and sentiment analysis (Radford et al., 2017).

In this paper, we connect these two lines of work and continue the trend of more general methods of transfer. We demonstrate language models can perform down-stream tasks in a zero-shot setting – without any parameter or architecture modification. We demonstrate this approach shows potential by highlighting the ability of language models to perform a wide range of tasks in a zero-shot setting. We achieve promising, competitive, and state of the art results depending on the task.

2. Approach

At the core of our approach is language modeling. Language modeling is usually framed as unsupervised distribution estimation from a set of examples $(x_1, x_2, ..., x_n)$ each composed of variable length sequences of symbols $(s_1, s_2, ..., s_n)$. Since language has a natural sequential ordering, it is common to factorize the joint probabilities over symbols as the product of conditional probabilities (Jelinek & Mercer, 1980) (Bengio et al., 2003):

$$p(x) = \prod_{i=1}^{n} p(s_n|s_1, ..., s_{n-1}) \quad (1)$$

This approach allows for tractable sampling from and estimation of $p(x)$ as well as any conditionals of the form $p(s_{n-k}, ..., s_n|s_1, ..., s_{n-k-1})$. In recent years, there have been significant improvements in the expressiveness of models that can compute these conditional probabilities, such as self-attention architectures like the Transformer (Vaswani et al., 2017).

Learning to perform a single task can be expressed in a probabilistic framework as estimating a conditional distribution $p(output|input)$. Since a general system should be able to perform many different tasks, even for the same input, it should condition not only on the input but also on the task to be performed. That is, it should model $p(output|input, task)$. This has been variously formalized in multitask and meta-learning settings. Task conditioning is often implemented at an architectural level, such as the task specific encoders and decoders in (Kaiser et al., 2017) or at an algorithmic level such as the inner and outer loop optimization framework of MAML (Finn et al., 2017). But as exemplified in McCann et al. (2018), language provides a flexible way to specify tasks, inputs, and outputs all as a sequence of symbols. For example, a translation training example can be written as the sequence `(translate to french, english text, french text)`. Likewise, a reading comprehension training example can be written as `(answer the question, document, question, answer)`. McCann et al. (2018) demonstrated it was possible to train a single model, the MQAN,

to infer and perform many different tasks on examples with this type of format.

Language modeling is also able to, in principle, learn the tasks of McCann et al. (2018) without the need for explicit supervision of which symbols are the outputs to be predicted. Since the supervised objective is the the same as the unsupervised objective but only evaluated on a subset of the sequence, the global minimum of the unsupervised objective is also the global minimum of the supervised objective. In this slightly toy setting, the concerns with density estimation as a principled training objective discussed in (Sutskever et al., 2015) are side stepped. The problem instead becomes whether we are able to, in practice, optimize the unsupervised objective to convergence. Preliminary experiments confirmed that sufficiently large language models are able to perform multitask learning in this toy-ish setup but learning is much slower than in explicitly supervised approaches.

While it is a large step from the well-posed setup described above to the messiness of "language in the wild", Weston (2016) argues, in the context of dialog, for the need to develop systems capable of learning from natural language directly and demonstrated a proof of concept – learning a QA task without a reward signal by using forward prediction of a teacher's outputs. While dialog is an attractive approach, we worry it is overly restrictive. The internet contains a vast amount of information that is passively available without the need for interactive communication. Our speculation is that a language model with sufficient capacity will begin to learn to infer and perform the tasks demonstrated in natural language sequences in order to better predict them, regardless of their method of procurement. If a language model is able to do this it will be, in effect, performing unsupervised multitask learning. We test whether this is the case by analyzing the performance of language models in a zero-shot setting on a wide variety of tasks.

2.1. Training Dataset

Most prior work trained language models on a single domain of text, such as news articles (Jozefowicz et al., 2016), Wikipedia (Merity et al., 2016), or fiction books (Kiros et al., 2015). Our approach motivates building as large and diverse a dataset as possible in order to collect natural language demonstrations of tasks in as varied of domains and contexts as possible.

A promising source of diverse and nearly unlimited text is web scrapes such as Common Crawl. While these archives are many orders of magnitude larger than current language modeling datasets, they have significant data quality issues. Trinh & Le (2018) used Common Crawl in their work on commonsense reasoning but noted a large amount of documents "whose content are mostly unintelligible". We observed similar data issues in our initial experiments with

"I'm not the cleverest man in the world, but like they say in French: **Je ne suis pas un imbecile [I'm not a fool].**
In a now-deleted post from Aug. 16, Soheil Eid, Tory candidate in the riding of Joliette, wrote in French: "**Mentez mentez, il en restera toujours quelque chose**," which translates as, "**Lie lie and something will always remain.**"
"I hate the word '**perfume**,"' Burr says. 'It's somewhat better in French: '**parfum**.'
If listened carefully at 29:55, a conversation can be heard between two guys in French: "**-Comment on fait pour aller de l'autre coté? -Quel autre coté?**", which means "**- How do you get to the other side? - What side?**".
If this sounds like a bit of a stretch, consider this question in French: **As-tu aller au cinéma?**, or **Did you go to the movies?**, which literally translates as Have-you to go to movies/theater?
"**Brevet Sans Garantie Du Gouvernement**", translated to English: "**Patented without government warranty**".

Table 1. Examples of naturally occurring demonstrations of English to French and French to English translation found throughout the WebText training set.

Common Crawl. Trinh & Le (2018)'s best results were achieved using a small subsample of Common Crawl which included only documents most similar to their target dataset, the Winograd Schema Challenge. While this is a pragmatic approach to improve performance on a specific task, we want to avoid making assumptions about the tasks to be performed ahead of time.

Instead, we created a new web scrape which emphasizes document quality. To do this we only scraped web pages which have been curated/filtered by humans. Manually filtering a full web scrape would be exceptionally expensive so as a starting point, we scraped all outbound links from Reddit, a social media platform, which received at least 3 karma. This can be thought of as a heuristic indicator for whether other users found the link interesting, educational, or just funny.

The resulting dataset, WebText, contains the text subset of these 45 million links. To extract the text from HTML responses we use a combination of the Dragnet (Peters & Lecocq, 2013) and Newspaper[1] content extractors. All results presented in this paper use a preliminary version of WebText which does not include links created after Dec 2017 and which after de-duplication and some heuristic based cleaning contains slightly over 8 million documents for a total of 40 GB of text. We removed all Wikipedia documents from WebText since it is a common data source for other datasets and could complicate analysis due to over-

[1] https://github.com/codelucas/newspaper

lapping training data with test evaluation tasks.

2.2. Input Representation

A general language model (LM) should be able to compute the probability of (and also generate) any string. Current large scale LMs include pre-processing steps such as lowercasing, tokenization, and out-of-vocabulary tokens which restrict the space of model-able strings. While processing Unicode strings as a sequence of UTF-8 bytes elegantly fulfills this requirement as exemplified in work such as Gillick et al. (2015), current byte-level LMs are not competitive with word-level LMs on large scale datasets such as the One Billion Word Benchmark (Al-Rfou et al., 2018). We observed a similar performance gap in our own attempts to train standard byte-level LMs on WebText.

Byte Pair Encoding (BPE) (Sennrich et al., 2015) is a practical middle ground between character and word level language modeling which effectively interpolates between word level inputs for frequent symbol sequences and character level inputs for infrequent symbol sequences. Despite its name, reference BPE implementations often operate on Unicode code points and not byte sequences. These implementations would require including the full space of Unicode symbols in order to model all Unicode strings. This would result in a base vocabulary of over 130,000 before any multi-symbol tokens are added. This is prohibitively large compared to the 32,000 to 64,000 token vocabularies often used with BPE. In contrast, a byte-level version of BPE only requires a base vocabulary of size 256. However, directly applying BPE to the byte sequence results in suboptimal merges due to BPE using a greedy frequency based heuristic for building the token vocabulary. We observed BPE including many versions of common words like `dog` since they occur in many variations such as `dog. dog! dog?` . This results in a sub-optimal allocation of limited vocabulary slots and model capacity. To avoid this, we prevent BPE from merging across character categories for any byte sequence. We add an exception for spaces which significantly improves the compression efficiency while adding only minimal fragmentation of words across multiple vocab tokens.

This input representation allows us to combine the empirical benefits of word-level LMs with the generality of byte-level approaches. Since our approach can assign a probability to any Unicode string, this allows us to evaluate our LMs on any dataset regardless of pre-processing, tokenization, or vocab size.

2.3. Model

We use a Transformer (Vaswani et al., 2017) based architecture for our LMs. The model largely follows the details of the OpenAI GPT model (Radford et al., 2018) with a few modifications. Layer normalization (Ba et al., 2016) was moved to the input of each sub-block, similar to a pre-activation residual network (He et al., 2016) and an additional layer normalization was added after the final self-attention block. A modified initialization which accounts for the accumulation on the residual path with model depth is used. We scale the weights of residual layers at initialization by a factor of $1/\sqrt{N}$ where N is the number of residual layers. The vocabulary is expanded to 50,257. We also increase the context size from 512 to 1024 tokens and a larger batchsize of 512 is used.

Parameters	Layers	d_{model}
117M	12	768
345M	24	1024
762M	36	1280
1542M	48	1600

Table 2. Architecture hyperparameters for the 4 model sizes.

3. Experiments

We trained and benchmarked four LMs with approximately log-uniformly spaced sizes. The architectures are summarized in Table 2. The smallest model is equivalent to the original GPT, and the second smallest equivalent to the largest model from BERT (Devlin et al., 2018). Our largest model, which we call GPT-2, has over an order of magnitude more parameters than GPT. The learning rate of each model was manually tuned for the best perplexity on a 5% held-out sample of WebText. All models still underfit WebText and held-out perplexity has as of yet improved given more training time.

3.1. Language Modeling

As an initial step towards zero-shot task transfer, we are interested in understanding how WebText LM's perform at zero-shot domain transfer on the primary task they are trained for – language modeling. Since our model operates on a byte level and does not require lossy pre-processing or tokenization, we can evaluate it on any language model benchmark. Results on language modeling datasets are commonly reported in a quantity which is a scaled or exponentiated version of the average negative log probability per canonical prediction unit - usually a character, a byte, or a word. We evaluate the same quantity by computing the log-probability of a dataset according to a WebText LM and dividing by the number of canonical units. For many of these datasets, WebText LMs would be tested significantly out-of-distribution, having to predict aggressively standardized text, tokenization artifacts such as disconnected punctuation and contractions, shuffled sentences, and even the string

	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14	46.54	0.99	1.08	18.3	**21.8**
117M	**35.13**	45.99	**87.65**	**83.4**	**29.41**	65.85	1.16	1.17	37.50	75.20
345M	**15.60**	55.48	**92.35**	**87.1**	**22.76**	47.33	1.01	**1.06**	26.37	55.72
762M	**10.87**	**60.12**	**93.45**	**88.0**	**19.93**	**40.31**	**0.97**	**1.02**	22.05	44.575
1542M	**8.63**	**63.24**	**93.30**	**89.05**	**18.34**	**35.76**	**0.93**	**0.98**	**17.48**	42.16

Table 3. Zero-shot results on many datasets. No training or fine-tuning was performed for any of these results. PTB and WikiText-2 results are from (Gong et al., 2018). CBT results are from (Bajgar et al., 2016). LAMBADA accuracy result is from (Hoang et al., 2018) and LAMBADA perplexity result is from (Grave et al., 2016). Other results are from (Dai et al., 2019).

<UNK> which is extremely rare in WebText - occurring only 26 times in 40 billion bytes. We report our main results in Table 3 using invertible de-tokenizers which remove as many of these tokenization / pre-processing artifacts as possible. Since these de-tokenizers are invertible, we can still calculate the log probability of a dataset and they can be thought of as a simple form of domain adaptation. We observe gains of 2.5 to 5 perplexity for GPT-2 with these de-tokenizers.

WebText LMs transfer well across domains and datasets, improving the state of the art on 7 out of the 8 datasets in a zero-shot setting. Large improvements are noticed on small datasets such as Penn Treebank and WikiText-2 which have only 1 to 2 million training tokens. Large improvements are also noticed on datasets created to measure long-term dependencies like LAMBADA (Paperno et al., 2016) and the Children's Book Test (Hill et al., 2015). Our model is still significantly worse than prior work on the One Billion Word Benchmark (Chelba et al., 2013). This is likely due to a combination of it being both the largest dataset and having some of the most destructive pre-processing - 1BW's sentence level shuffling removes all long-range structure.

3.2. Children's Book Test

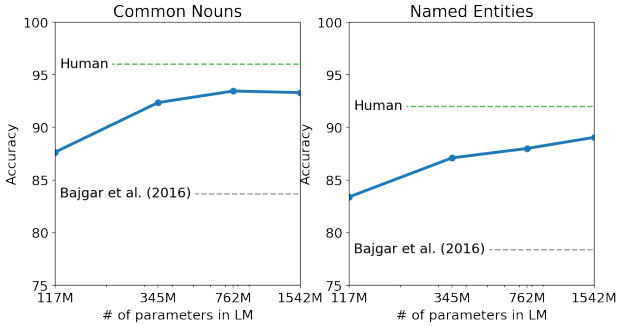


Figure 2. Performance on the Children's Book Test as a function of model capacity. Human performance are from Bajgar et al. (2016), instead of the much lower estimates from the original paper.

The Children's Book Test (CBT) (Hill et al., 2015) was created to examine the performance of LMs on different categories of words: named entities, nouns, verbs, and prepositions. Rather than reporting perplexity as an evaluation metric, CBT reports accuracy on an automatically constructed cloze test where the task is to predict which of 10 possible choices for an omitted word is correct. Following the LM approach introduced in the original paper, we compute the probability of each choice and the rest of the sentence conditioned on this choice according to the LM, and predict the one with the highest probability. As seen in Figure 2 performance steadily improves as model size is increased and closes the majority of the gap to human performance on this test. Data overlap analysis showed one of the CBT test set books, The Jungle Book by Rudyard Kipling, is in WebText, so we report results on the validation set which has no significant overlap. GPT-2 achieves new state of the art results of 93.3% on common nouns and 89.1% on named entities. A de-tokenizer was applied to remove PTB style tokenization artifacts from CBT.

3.3. LAMBADA

The LAMBADA dataset (Paperno et al., 2016) tests the ability of systems to model long-range dependencies in text. The task is to predict the final word of sentences which require at least 50 tokens of context for a human to successfully predict. GPT-2 improves the state of the art from 99.8 (Grave et al., 2016) to 8.6 perplexity and increases the accuracy of LMs on this test from 19% (Dehghani et al., 2018) to 52.66%. Investigating GPT-2's errors showed most predictions are valid continuations of the sentence, but are not valid final words. This suggests that the LM is not using the additional useful constraint that the word must be the final of the sentence. Adding a stop-word filter as an approximation to this further increases accuracy to 63.24%, improving the overall state of the art on this task by 4%. The previous state of the art (Hoang et al., 2018) used a different restricted prediction setting where the outputs of the model were constrained to only words that appeared in the context. For GPT-2, this restriction is harmful rather than helpful

since 19% of answers are not in context. We use a version of the dataset without preprocessing.

3.4. Winograd Schema Challenge

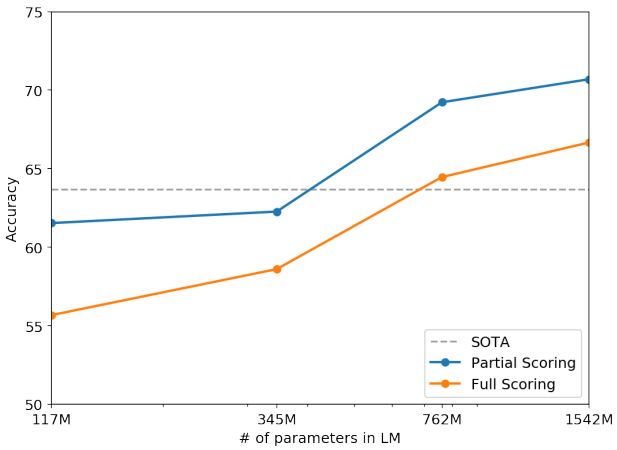


Figure 3. Performance on the Winograd Schema Challenge as a function of model capacity.

The Winograd Schema challenge (Levesque et al., 2012) was constructed to measure the capability of a system to perform commonsense reasoning by measuring its ability to resolve ambiguities in text. Recently Trinh & Le (2018) demonstrated significant progress on this challenge using LMs, by predicting the resolution of the ambiguity with higher probability. We follow their problem formulation and visualize the performance of our models with both full and partial scoring techniques in Figure 3. GPT-2 improves state of the art accuracy by 7%, achieving 70.70%. The dataset is quite small with only 273 examples so we recommend reading Trichelair et al. (2018) to help contextualize this result.

3.5. Reading Comprehension

The Conversation Question Answering dataset (CoQA) Reddy et al. (2018) consists of documents from 7 different domains paired with natural language dialogues between a question asker and a question answerer about the document. CoQA tests reading comprehension capabilities and also the ability of models to answer questions that depend on conversation history (such as "Why?").

Greedy decoding from GPT-2 when conditioned on a document, the history of the associated conversation, and a final token `A:` achieves 55 F1 on the development set. This matches or exceeds the performance of 3 out of 4 baseline systems without using the 127,000+ manually collected question answer pairs those baselines were trained on. The supervised SOTA, a BERT based system (Devlin et al., 2018), is nearing the 89 F1 performance of humans. While GPT-2's performance is exciting for a system without any supervised training, some inspection of its answers and errors suggests GPT-2 often uses simple retrieval based heuristics such as *answer with a name from the document in response to a who question*.

	R-1	R-2	R-L	R-AVG
Bottom-Up Sum	**41.22**	**18.68**	**38.34**	**32.75**
Lede-3	40.38	17.66	36.62	31.55
Seq2Seq + Attn	31.33	11.81	28.83	23.99
GPT-2 `TL;DR:`	29.34	8.27	26.58	21.40
Random-3	28.78	8.63	25.52	20.98
GPT-2 no hint	21.58	4.03	19.47	15.03

Table 4. Summarization performance as measured by ROUGE F1 metrics on the CNN and Daily Mail dataset. Bottom-Up Sum is the SOTA model from (Gehrmann et al., 2018)

3.6. Summarization

We test GPT-2's ability to perform summarization on the CNN and Daily Mail dataset (Nallapati et al., 2016). To induce summarization behavior we add the text `TL;DR:` after the article and generate 100 tokens with Top-k random sampling (Fan et al., 2018) with $k = 2$ which reduces repetition and encourages more abstractive summaries than greedy decoding. We use the first 3 generated sentences in these 100 tokens as the summary. While qualitatively the generations resemble summaries, as shown in Table 14, they often focus on recent content from the article or confuse specific details such as how many cars were involved in a crash or whether a logo was on a hat or shirt. On the commonly reported ROUGE 1,2,L metrics the generated summaries only begin to approach the performance of classic neural baselines and just barely outperforms selecting 3 random sentences from the article. GPT-2's performance drops by 6.4 points on the aggregate metric when the task hint is removed which demonstrates the ability to invoke task specific behavior in a language model with natural language.

3.7. Translation

We test whether GPT-2 has begun to learn how to translate from one language to another. In order to help it infer that this is the desired task, we condition the language model on a context of example pairs of the format `english sentence = french sentence` and then after a final prompt of `english sentence =` we sample from the model with greedy decoding and use the first generated sentence as the translation. On the WMT-14 English-French test set, GPT-2 gets 5 BLEU, which is slightly worse than a word-by-word substitution with a bilingual lexicon inferred in previous work on unsupervised word translation

Question	Generated Answer	Correct	Probability
Who wrote the book the origin of species?	Charles Darwin	✓	83.4%
Who is the founder of the ubuntu project?	Mark Shuttleworth	✓	82.0%
Who is the quarterback for the green bay packers?	Aaron Rodgers	✓	81.1%
Panda is a national animal of which country?	China	✓	76.8%
Who came up with the theory of relativity?	Albert Einstein	✓	76.4%
When was the first star wars film released?	1977	✓	71.4%
What is the most common blood type in sweden?	A	✗	70.6%
Who is regarded as the founder of psychoanalysis?	Sigmund Freud	✓	69.3%
Who took the first steps on the moon in 1969?	Neil Armstrong	✓	66.8%
Who is the largest supermarket chain in the uk?	Tesco	✓	65.3%
What is the meaning of shalom in english?	peace	✓	64.0%
Who was the author of the art of war?	Sun Tzu	✓	59.6%
Largest state in the us by land mass?	California	✗	59.2%
Green algae is an example of which type of reproduction?	parthenogenesis	✗	56.5%
Vikram samvat calender is official in which country?	India	✓	55.6%
Who is mostly responsible for writing the declaration of independence?	Thomas Jefferson	✓	53.3%
What us state forms the western boundary of montana?	Montana	✗	52.3%
Who plays ser davos in game of thrones?	Peter Dinklage	✗	52.1%
Who appoints the chair of the federal reserve system?	Janet Yellen	✗	51.5%
State the process that divides one nucleus into two genetically identical nuclei?	mitosis	✓	50.7%
Who won the most mvp awards in the nba?	Michael Jordan	✗	50.2%
What river is associated with the city of rome?	the Tiber	✓	48.6%
Who is the first president to be impeached?	Andrew Johnson	✓	48.3%
Who is the head of the department of homeland security 2017?	John Kelly	✓	47.0%
What is the name given to the common currency to the european union?	Euro	✓	46.8%
What was the emperor name in star wars?	Palpatine	✓	46.5%
Do you have to have a gun permit to shoot at a range?	No	✓	46.4%
Who proposed evolution in 1859 as the basis of biological development?	Charles Darwin	✓	45.7%
Nuclear power plant that blew up in russia?	Chernobyl	✓	45.7%
Who played john connor in the original terminator?	Arnold Schwarzenegger	✗	45.2%

Table 5. The 30 most confident answers generated by GPT-2 on the development set of Natural Questions sorted by their probability according to GPT-2. None of these questions appear in WebText according to the procedure described in Section 4.

(Conneau et al., 2017b). On the WMT-14 French-English test set, GPT-2 is able to leverage its very strong English language model to perform significantly better, achieving 11.5 BLEU. This outperforms several unsupervised machine translation baselines from (Artetxe et al., 2017) and (Lample et al., 2017) but is still much worse than the 33.5 BLEU of the current best unsupervised machine translation approach (Artetxe et al., 2019). Performance on this task was surprising to us, since we deliberately removed non-English webpages from WebText as a filtering step. In order to confirm this, we ran a byte-level language detector[2] on WebText which detected only 10MB of data in the French language which is approximately 500x smaller than the monolingual French corpus common in prior unsupervised machine translation research.

3.8. Question Answering

A potential way to test what information is contained within a language model is to evaluate how often it generates the correct answer to factoid-style questions. Previous showcasing of this behavior in neural systems where all information is stored in parameters such as *A Neural Conversational Model* (Vinyals & Le, 2015) reported qualitative results due to the lack of high-quality evaluation datasets. The recently introduced Natural Questions dataset (Kwiatkowski et al., 2019) is a promising resource to test this more quantitatively. Similar to translation, the context of the language model is seeded with example question answer pairs which helps the model infer the short answer style of the dataset. GPT-2 answers 4.1% of questions correctly when evaluated by the exact match metric commonly used on reading comprehension datasets like SQUAD.[3] As a comparison point, the smallest model does not exceed the 1.0% accuracy of an incredibly simple baseline which returns the most common answer for each question type (who, what, where, etc...). GPT-2 answers 5.3 times more questions correctly, suggesting that model capacity has been a major factor in the poor performance of neural systems on this kind of task as of yet. The probability GPT-2 assigns to its generated answers is well calibrated and GPT-2 has an accuracy of 63.1% on the 1% of questions it is most confident in. The 30 most confident answers generated by GPT-2 on development set questions are shown in Table 5. The performance of GPT-2 is still much, much, worse than the 30 to 50% range of open domain question answering systems which hybridize information retrieval with extractive document question answering (Alberti et al., 2019).

[2] https://github.com/CLD2Owners/cld2

[3] Alec, who previously thought of himself as good at random trivia, answered 17 of 100 randomly sampled examples correctly when tested in the same setting as GPT-2. He actually only got 14 right but he should have gotten those other 3

	PTB	WikiText-2	enwik8	text8	Wikitext-103	1BW
Dataset train	**2.67%**	0.66%	**7.50%**	2.34%	**9.09%**	**13.19%**
WebText train	0.88%	**1.63%**	6.31%	**3.94%**	2.42%	3.75%

Table 6. Percentage of test set 8 grams overlapping with training sets.

4. Generalization vs Memorization

Recent work in computer vision has shown that common image datasets contain a non-trivial amount of near-duplicate images. For instance CIFAR-10 has 3.3% overlap between train and test images (Barz & Denzler, 2019). This results in an over-reporting of the generalization performance of machine learning systems. As the size of datasets increases this issue becomes increasingly likely which suggests a similar phenomena could be happening with WebText. Therefore it is important to analyze how much test data also shows up in the training data.

To study this we created Bloom filters containing 8-grams of WebText training set tokens. To improve recall, strings were normalized to contain only lower-cased alphanumeric words with a single space as a delimiter. The Bloom filters were constructed such that the false positive rate is upper bounded by $\frac{1}{10^8}$. We further verified the low false positive rate by generating 1M strings, of which zero were found by the filter.

These Bloom filters let us calculate, given a dataset, the percentage of 8-grams from that dataset that are also found in the WebText training set. Table 6 shows this overlap analysis for the test sets of common LM benchmarks. Common LM datasets' test sets have between 1-6% overlap with WebText train, with an average of overlap of 3.2%. Somewhat surprisingly, many datasets have larger overlaps with their own training splits, with an average of 5.9% overlap.

Our approach optimizes for recall, and while manual inspection of the overlaps shows many common phrases, there are many longer matches that are due to duplicated data. This is not unique to WebText. For instance, we discovered that the test set of WikiText-103 has an article which is also in the training dataset. Since there are only 60 articles in the test set there is at least an overlap of 1.6%.[4] Potentially more worryingly, 1BW has an overlap of nearly 13.2% with its own training set according to our procedure.

For the Winograd Schema Challenge, we found only 10 schemata which had any 8-gram overlaps with the WebText training set. Of these, 2 were spurious matches. Of the remaining 8, only 1 schema appeared in any contexts that gave away the answer.

For CoQA, about 15% of documents in the news domain are already in WebText and the model performs about 3 F1 better on these. CoQA's development set metric reports the average performance over 5 different domains and we measure a gain of about 0.5-1.0 F1 due to overlap across the various domains. However, no actual training questions or answers are in WebText since CoQA was released after the cutoff date for links in WebText.

On LAMBADA, the average overlap is 1.2%. GPT-2 performs about 2 perplexity better on examples with greater than 15% overlap. Recalculating metrics when excluding all examples with any overlap shifts results from 8.6 to 8.7 perplexity and reduces accuracy from 63.2% to 62.9%. This very small change in overall results is likely due to only 1 in 200 examples having significant overlap.

Overall, our analysis suggests that data overlap between WebText training data and specific evaluation datasets provides a small but consistent benefit to reported results. However, for most datasets we do not notice significantly larger overlaps than those already existing between standard training and test sets, as Table 6 highlights.

Understanding and quantifying how highly similar text impacts performance is an important research question. Better de-duplication techniques such as scalable fuzzy matching could also help better answer these questions. For now, we recommend the use of n-gram overlap based de-duplication as an important verification step and sanity check during the creation of training and test splits for new NLP datasets.

Another potential way of determining whether the performance of WebText LMs is attributable to memorization is inspecting their performance on their own held-out set. As shown in Figure 4, performance on both the training and test sets of WebText are similar and improve together as model size is increased. This suggests even GPT-2 is still underfitting on WebText in many ways.

GPT-2 is also able to write news articles about the discovery of talking unicorns. An example is provided in Table 13.

5. Related Work

A significant portion of this work measured the performance of larger language models trained on larger datasets. This

[4] A significant portion of additional overlap is due to editors reusing some paragraphs across multiple articles with a shared theme such as various battles in the Korean War.

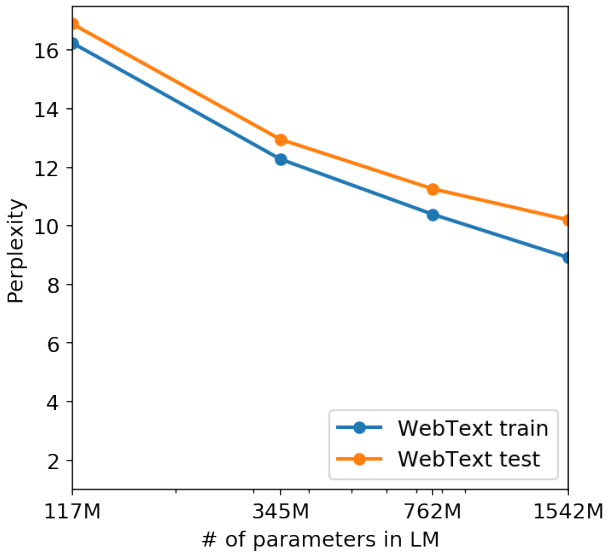


Figure 4. The performance of LMs trained on WebText as a function of model size.

is similar to the work of Jozefowicz et al. (2016) which scaled RNN based language models on the 1 Billion Word Benchmark. Bajgar et al. (2016) also previously improved results on the Children's Book Test by creating a much larger training dataset out of Project Gutenberg to supplement the standard training dataset. Hestness et al. (2017) conducted a thorough analysis of how the performance of various deep learning models changes as a function of both model capacity and dataset size. Our experiments, while much noisier across tasks, suggest similar trends hold for sub-tasks of an objective and continue into the 1B+ parameter regime.

Interesting learned functionality in generative models has been documented before such as the cells in an RNN language model performing line-width tracking and quote/comment detection Karpathy et al. (2015). More inspirational to our work was the observation of Liu et al. (2018) that a model trained to generate Wikipedia articles also learned to translate names between languages.

Previous work has explored alternative approaches to filtering and constructing a large text corpus of web pages, such as the iWeb Corpus (Davies, 2018).

There has been extensive work on pre-training methods for language tasks. In addition to those mentioned in the introduction, GloVe (Pennington et al., 2014) scaled word vector representation learning to all of Common Crawl. An influential early work on deep representation learning for text was *Skip-thought Vectors* (Kiros et al., 2015). McCann et al. (2017) explored the use of representations derived from machine translation models and Howard & Ruder (2018) improved the RNN based fine-tuning approaches of (Dai & Le, 2015). (Conneau et al., 2017a) studied the transfer performance of representations learned by natural language inference models and (Subramanian et al., 2018) explored large-scale multitask training.

(Ramachandran et al., 2016) demonstrated that seq2seq models benefit from being initialized with pre-trained language models as encoders and decoders. More recent work has shown that LM pre-training is helpful when fine-tuned for difficult generation tasks like chit-chat dialog and dialog based question answering systems as well (Wolf et al., 2019) (Dinan et al., 2018).

6. Discussion

Much research has been dedicated to learning (Hill et al., 2016), understanding (Levy & Goldberg, 2014), and critically evaluating (Wieting & Kiela, 2019) the representations of both supervised and unsupervised pre-training methods. Our results suggest that unsupervised task learning is an additional promising area of research to explore. These findings potentially help explain the widespread success of pre-training techniques for down-stream NLP tasks as we show that, in the limit, one of these pre-training techniques begins to learn to perform tasks directly without the need for supervised adaption or modification.

On reading comprehension the performance of GPT-2 is competitive with supervised baselines in a zero-shot setting. However, on other tasks such as summarization, while it is qualitatively performing the task, its performance is still only rudimentary according to quantitative metrics. While suggestive as a research result, in terms of practical applications, the zero-shot performance of GPT-2 is still far from use-able.

We have studied the zero-shot performance of WebText LMs on many canonical NLP tasks, but there are many additional tasks that could be evaluated. There are undoubtedly many practical tasks where the performance of GPT-2 is still no better than random. Even on common tasks that we evaluated on, such as question answering and translation, language models only begin to outperform trivial baselines when they have sufficient capacity.

While zero-shot performance establishes a baseline of the potential performance of GPT-2 on many tasks, it is not clear where the ceiling is with finetuning. On some tasks, GPT-2's fully abstractive output is a significant departure from the extractive pointer network (Vinyals et al., 2015) based outputs which are currently state of the art on many question answering and reading comprehension datasets. Given the prior success of fine-tuning GPT, we plan to investigate fine-tuning on benchmarks such as decaNLP and GLUE, especially since it is unclear whether the additional

training data and capacity of GPT-2 is sufficient to overcome the inefficiencies of uni-directional representations demonstrated by BERT (Devlin et al., 2018).

7. Conclusion

When a large language model is trained on a sufficiently large and diverse dataset it is able to perform well across many domains and datasets. GPT-2 zero-shots to state of the art performance on 7 out of 8 tested language modeling datasets. The diversity of tasks the model is able to perform in a zero-shot setting suggests that high-capacity models trained to maximize the likelihood of a sufficiently varied text corpus begin to learn how to perform a surprising amount of tasks without the need for explicit supervision.[5]

Acknowledgements

Thanks to everyone who wrote the text, shared the links, and upvoted the content in WebText. Many millions of people were involved in creating the data that GPT-2 was trained on. Also thanks to all the Googlers who helped us with training infrastructure, including Zak Stone, JS Riehl, Jonathan Hseu, Russell Power, Youlong Cheng, Noam Shazeer, Solomon Boulos, Michael Banfield, Aman Gupta, Daniel Sohn, and many more. Finally thanks to the people who gave feedback on drafts of the paper: Jacob Steinhardt, Sam Bowman, Geoffrey Irving, and Madison May.

References

Al-Rfou, R., Choe, D., Constant, N., Guo, M., and Jones, L. Character-level language modeling with deeper self-attention. *arXiv preprint arXiv:1808.04444*, 2018.

Alberti, C., Lee, K., and Collins, M. A bert baseline for the natural questions. *arXiv preprint arXiv:1901.08634*, 2019.

Alcorn, M. A., Li, Q., Gong, Z., Wang, C., Mai, L., Ku, W.-S., and Nguyen, A. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. *arXiv preprint arXiv:1811.11553*, 2018.

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International Conference on Machine Learning*, pp. 173–182, 2016.

Artetxe, M., Labaka, G., Agirre, E., and Cho, K. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*, 2017.

Artetxe, M., Labaka, G., and Agirre, E. An effective approach to unsupervised machine translation. *arXiv preprint arXiv:1902.01313*, 2019.

[5]Preliminary code for downloading and using the small model is available at https://github.com/openai/gpt-2

Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Bajgar, O., Kadlec, R., and Kleindienst, J. Embracing data abundance: Booktest dataset for reading comprehension. *arXiv preprint arXiv:1610.00956*, 2016.

Barz, B. and Denzler, J. Do we train on test data? purging cifar of near-duplicates. *arXiv preprint arXiv:1902.00423*, 2019.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

Bowman, S. R., Pavlick, E., Grave, E., Van Durme, B., Wang, A., Hula, J., Xia, P., Pappagari, R., McCoy, R. T., Patel, R., et al. Looking for elmo's friends: Sentence-level pretraining beyond language modeling. *arXiv preprint arXiv:1812.10860*, 2018.

Caruana, R. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., and Robinson, T. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*, 2013.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537, 2011.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*, 2017a.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017b.

Dai, A. M. and Le, Q. V. Semi-supervised sequence learning. In *Advances in neural information processing systems*, pp. 3079–3087, 2015.

Dai, Z., Yang, Z., Yang, Y., Cohen, W. W., Carbonell, J., Le, Q. V., and Salakhutdinov, R. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

Davies, M. The 14 billion word iweb corpus. *https://corpus.byu.edu/iWeb/*, 2018.

Dehghani, M., Gouws, S., Vinyals, O., Uszkoreit, J., and Kaiser, Ł. Universal transformers. *arXiv preprint arXiv:1807.03819*, 2018.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., and Weston, J. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*, 2018.

Fan, A., Lewis, M., and Dauphin, Y. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*, 2018.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. *arXiv preprint arXiv:1703.03400*, 2017.

Gehrmann, S., Deng, Y., and Rush, A. M. Bottom-up abstractive summarization. *arXiv preprint arXiv:1808.10792*, 2018.

Gillick, D., Brunk, C., Vinyals, O., and Subramanya, A. Multilingual language processing from bytes. *arXiv preprint arXiv:1512.00103*, 2015.

Gong, C., He, D., Tan, X., Qin, T., Wang, L., and Liu, T.-Y. Frage: frequency-agnostic word representation. In *Advances in Neural Information Processing Systems*, pp. 1341–1352, 2018.

Grave, E., Joulin, A., and Usunier, N. Improving neural language models with a continuous cache. *arXiv preprint arXiv:1612.04426*, 2016.

He, K., Zhang, X., Ren, S., and Sun, J. Identity mappings in deep residual networks. In *European conference on computer vision*, pp. 630–645. Springer, 2016.

Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M., Ali, M., Yang, Y., and Zhou, Y. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

Hill, F., Bordes, A., Chopra, S., and Weston, J. The goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301*, 2015.

Hill, F., Cho, K., and Korhonen, A. Learning distributed representations of sentences from unlabelled data. *arXiv preprint arXiv:1602.03483*, 2016.

Hoang, L., Wiseman, S., and Rush, A. M. Entity tracking improves cloze-style reading comprehension. *arXiv preprint arXiv:1810.02891*, 2018.

Howard, J. and Ruder, S. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pp. 328–339, 2018.

Jelinek, F. and Mercer, R. L. Interpolated estimation of markov source parameters from sparse data. *In Proceedings of the Workshop on Pattern Recognition in Practice, Amsterdam, The Netherlands: North-Holland, May.*, 1980.

Jia, R. and Liang, P. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*, 2017.

Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., and Wu, Y. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*, 2016.

Kaiser, L., Gomez, A. N., Shazeer, N., Vaswani, A., Parmar, N., Jones, L., and Uszkoreit, J. One model to learn them all. *arXiv preprint arXiv:1706.05137*, 2017.

Karpathy, A., Johnson, J., and Fei-Fei, L. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*, 2015.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, pp. 201611835, 2017.

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., and Fidler, S. Skip-thought vectors. In *Advances in neural information processing systems*, pp. 3294–3302, 2015.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Kwiatkowski, T., Palomaki, J., Rhinehart, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Kelcey, M., Devlin, J., et al. Natural questions: a benchmark for question answering research. 2019.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, 2017.

Lample, G., Conneau, A., Denoyer, L., and Ranzato, M. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*, 2017.

Levesque, H., Davis, E., and Morgenstern, L. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*, 2012.

Levy, O. and Goldberg, Y. Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, pp. 2177–2185, 2014.

Liu, P. J., Saleh, M., Pot, E., Goodrich, B., Sepassi, R., Kaiser, L., and Shazeer, N. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*, 2018.

McCann, B., Bradbury, J., Xiong, C., and Socher, R. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems*, pp. 6294–6305, 2017.

McCann, B., Keskar, N. S., Xiong, C., and Socher, R. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*, 2018.

Merity, S., Xiong, C., Bradbury, J., and Socher, R. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013.

Nallapati, R., Zhou, B., Gulcehre, C., Xiang, B., et al. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*, 2016.

Paperno, D., Kruszewski, G., Lazaridou, A., Pham, Q. N., Bernardi, R., Pezzelle, S., Baroni, M., Boleda, G., and Fernández, R. The lambada dataset: Word prediction requiring a broad discourse context. *arXiv preprint arXiv:1606.06031*, 2016.

Pennington, J., Socher, R., and Manning, C. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

Peters, M. E. and Lecocq, D. Content extraction using diverse feature sets. In *Proceedings of the 22nd International Conference on World Wide Web*, pp. 89–90. ACM, 2013.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

Radford, A., Jozefowicz, R., and Sutskever, I. Learning to generate reviews and discovering sentiment. *arXiv preprint arXiv:1704.01444*, 2017.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. Improving language understanding by generative pre-training. 2018.

Ramachandran, P., Liu, P. J., and Le, Q. V. Unsupervised pretraining for sequence to sequence learning. *arXiv preprint arXiv:1611.02683*, 2016.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do cifar-10 classifiers generalize to cifar-10? *arXiv preprint arXiv:1806.00451*, 2018.

Reddy, S., Chen, D., and Manning, C. D. Coqa: A conversational question answering challenge. *arXiv preprint arXiv:1808.07042*, 2018.

Schwartz, R., Sap, M., Konstas, I., Zilles, L., Choi, Y., and Smith, N. A. Story cloze task: Uw nlp system. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pp. 52–55, 2017.

See, A., Liu, P. J., and Manning, C. D. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*, 2017.

Sennrich, R., Haddow, B., and Birch, A. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

Subramanian, S., Trischler, A., Bengio, Y., and Pal, C. J. Learning general purpose distributed sentence representations via large scale multi-task learning. *arXiv preprint arXiv:1804.00079*, 2018.

Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pp. 3104–3112, 2014.

Sutskever, I., Jozefowicz, R., Gregor, K., Rezende, D., Lillicrap, T., and Vinyals, O. Towards principled unsupervised learning. *arXiv preprint arXiv:1511.06440*, 2015.

Trichelair, P., Emami, A., Cheung, J. C. K., Trischler, A., Suleman, K., and Diaz, F. On the evaluation of common-sense reasoning in natural language understanding. *arXiv preprint arXiv:1811.01778*, 2018.

Trinh, T. H. and Le, Q. V. A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*, 2018.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

Vinyals, O. and Le, Q. A neural conversational model. *arXiv preprint arXiv:1506.05869*, 2015.

Vinyals, O., Fortunato, M., and Jaitly, N. Pointer networks. In *Advances in Neural Information Processing Systems*, pp. 2692–2700, 2015.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

Weston, J. E. Dialog-based language learning. In *Advances in Neural Information Processing Systems*, pp. 829–837, 2016.

Wieting, J. and Kiela, D. No training required: Exploring random encoders for sentence classification. *arXiv preprint arXiv:1901.10444*, 2019.

Wolf, T., Sanh, V., Chaumond, J., and Delangue, C. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*, 2019.

Yogatama, D., d'Autume, C. d. M., Connor, J., Kocisky, T., Chrzanowski, M., Kong, L., Lazaridou, A., Ling, W., Yu, L., Dyer, C., et al. Learning and evaluating general linguistic intelligence. *arXiv preprint arXiv:1901.11373*, 2019.

8. Appendix A: Samples

8.1. Model capacity

To complement the reported perplexity gains of bigger LMs on WebText show in Figure 4, Tables 7 through 11 show side-by-side completions of the smallest WebText LM and GPT-2 on random unseen WebText test set articles.

8.2. Text Memorization

We observe some memorizing behavior in GPT-2 on longer strings that are repeated many times in the dataset such as famous quotes or speeches. For example, when conditioned on the first sentence and a half of the Gettysburg Address (which occurs approximately 40 times throughout WebText), an argmax decode from GPT-2 recovers the speech. Even when sampling without truncation, we find that the model copies the speech for awhile before drifting, albeit in a similar style. It typically drifts within 100-200 tokens, and displays widening diversity once it drifts.

To quantify how often exact memorization shows up in samples, we generated samples from GPT-2 conditioned on WebText test set articles and compared the overlap rates of GPT-2's generations to the overlap rates of the ground-truth completions. The results of this analysis are shown below and suggest that GPT-2 repeats text from the training set less often then the baseline rate of held-out articles.

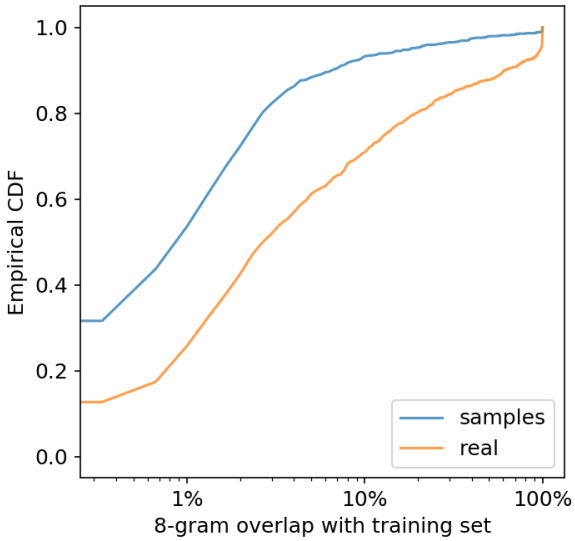


Figure 5. CDF of percentage 8-gram overlap with WebText training set, for both WebText test set and samples (conditioned on WebText test set, with top-k truncated random sampling with $k = 40$). Most samples have less than 1% overlap, including over 30% of samples with no overlap, whereas the median for test set is 2.6% overlap.

8.3. Diversity

Table 12 shows multiple completions of the same random WebText test set context, showing the diversity of completions with standard sampling settings.

8.4. Robustness

Table 13 shows the previously mentioned talking unicorns news article. We find the model to be capable of handling out of distribution contexts, but the quality of these samples is generally lower.

<table>
<tr><th colspan="2">Context (WebText test)</th></tr>
<tr><td colspan="2">Outfit 1: Typical

This pairing was the first outfit I thought of when I bought the shoes. It's like a summer version of this Jake Grantham outfit; in fact, my shoes are close to the colors of his Nike Racers! Instead of a heavy Harris Tweed jacket and denim shirt, I'm wearing a cotton DB jacket and and a linen shirt. Both fabrics (in these colors) are an absolute must for summer, as they go with both dark and and light pants! As you can see, they pair wonderfully with the dark jeans and shoes. It's a pseudo menswear/prep outfit.

Overall, this is a very casual outfit which is why I paired my sneakers with it. I'm not about wearing a full wool suit with sneakers (as GQ shows a lot) but I'm definitely open to keeping things casual, like this cotton DB. Casual fabrics are key to pulling off your sneakers in a dressed down menswear outfit. I'd even suggest to wear these sneakers with a khaki chino suit or a white linen suit. Just be sure to ditch the tie or wear a tee or polo; wearing a tie with sneakers is a bit too much</td></tr>
<tr><th>Small model completion</th><th>GPT-2 completion</th></tr>
<tr><td>.

Outfit 2: High End

This outfit was just in the eye of the storm. It's a hybrid of what ASICS does with streetwear (like J. Crew) and what J. Crew does with casual pieces that are a bit more traditional. Again, as I mentioned before, it's a low budget item and they can be very expensive. These will definitely be an issue if you're trying to work as a casual sneakermaker, and that's OK but don't hesitate if you want to try the lower price versions of their outfits!

This version features a very low cut-cut neckline (which is not a bad thing), slightly longer legs, and a pair of long legs to keep things nice. My preferred color of color these items are blue-gray (which is my first colors, as they are a light white), or black (which is my favorite of my designs). As always, my pick of colors from this store have nothing to do with that other than the design. For a high end, I would recommend this to everyone!

Outfit 3: High Light

This was a "high light" look while I had a chance to get some color. I think the color is based on</td><td>for me!

Outfit 2: Casual Jacket

I know that some people might be opposed to wearing sneakers in a turtleneck, but I wanted to be true to myself, so I went with a slim fitting turtleneck in a color more similar to my favorite color of the day. Overall, it's a classic turtleneck, with nothing too flashy. I did swap my shoes for these ASICS Tiger Killshots and I wish I hadn't did that. The shoe itself is very comfortable, however, I found that wearing them in these sneakers made my feet look a bit larger. I'd wear them with denim, chinos, whatever, but would never recommend wearing them alone. There are just too many variables involved.

Outfit 3: Dress

This is the final outfit I wore the ASICS Tiger Killshots. This version of the ASICS is a much more modern and streamlined version. The shoe has the most modern and minimal design that I can think of! With a lot of black, these sneakers look very cool. They'll definitely be very comfortable for summer.

To me, wearing them with a suit is the easiest way to have a good pair of sneaker shoes. That's pretty much it! If</td></tr>
</table>

Table 7. Random unseen contexts (top), and non-cherry-picked completions from both the smallest (left) and largest (right) models. Contexts are 768 tokens, with approximately 256 tokens worth of paragraphs shown. Completions are 256 tokens and fully shown. Top-k random sampling with $k = 40$ was used for generation.

Context (WebText test)	
Some of the most glorious historical attractions in Spain date from the period of Muslim rule, including The Mezquita, built as the Great Mosque of Córdoba and the Medina Azahara, also in Córdoba and now in ruins but still visitable as such and built as the Madinat al-Zahra, the Palace of al-Andalus; and the Alhambra in Granada, a splendid, intact palace. There are also two synagogues still standing that were built during the era of Muslim Spain: Santa María la Blanca in Toledo and the Synagogue of Córdoba, in the Old City. Reconquista and Imperial era [edit]	
Small model completion	**GPT-2 completion**
The era of Muslim Spain is very different from that of other parts of the world, with a rather extensive medieval history. While the Spanish Empire existed during the second half of the 9th century BC it was quite slowly transformed to a small, secularized empire by 1150 A.D., in a much more prosperous and diverse state. In 1050 CE the emperor of Naples, at least partially in characterised by a series of wars and invasions from the north-western coast to the western coast of the region (e.g., on the Danube River which extended all the way from Gibraltar into the south of Spain, but was quickly destroyed on the same day by the Roman army), assumed the title of Emperor of Naples. He made a large contribution over the last few decades to restore and maintain the empire, the first of which to the west. His successors were Ferdinand, who took power from the Papal Throne in 1148, while the first imperial successor was King Alfred II of Sweden. The empire was again made one of the few independent states in the European Continent and although no more than 200 million people lived on its territory during this period the people had much to say to the Emperor. From 1720 on the Muslims made up half the population of Spain.	The Reconquista (literally, ”The return to the country”) of 1092 and 1095 saw the establishment in the Iberian peninsula of a united kingdom of Castile and Aragon in which a Muslim dynasty would exist. This was the last time in Europe where Muslim rule would be considered ”native” and the last occasion where Christian-inflamed and organized conflicts would make themselves known in Europe. The reign of Muslim Moors in Spain is often referred to, especially among non-Muslims, as the ”Conquest of Granada” and Spain is sometimes referred to as Spain in ”Granada”. The last years of Muslim rule were followed by the Spanish Reconquista where the Muslims were driven or compelled out of the Iberian Peninsula with the victory of Ferdinand IX. in the Battle of Poitiers in March 1189 and the conquest of Jerusalem in 1187 by Saladin. The Reconquista period in Spain came to an end when the Moors retreated north. However the Moorish rule would not be fully replaced by the Christian rule until the Reconquista era resumed when the Reconquista lasted from 1229 to 1492. The last Muslim state in Spain, Granada, was conquered by the

Table 8. Random unseen contexts (top), and non-cherry-picked completions from both the smallest (left) and largest (right) models. Contexts are 768 tokens, with approximately 256 tokens worth of paragraphs shown. Completions are 256 tokens and fully shown. Top-k random sampling with $k = 40$ was used for generation.

Context (WebText test)	
Pinky The pink ghost's AI is designed to "feel" opposite of the red ghost's behavior. Pinky actually attempts to get out in front of Pac-Man. This is accomplished by setting the target 4 tiles ahead of Pac-Man's current location in the direction that Pac-Man is travelling. One exception to this is when Pac-Man is traveling up. Due to an overflow bug in the code, the calculation includes a left offset equal to the expected up offset. The use of Pac-Man's expected position allows the player to exploit the navigation code for Pinky. This was most often seen in situations where Pac-Man and Pinky are in close proximity to each other. For example, if they were headed toward one another, once they are less than 4 tiles apart, the target tile is now behind Pinky. If he encounters an intersection while in this configuration, he will attempt to double-back (by turning) in order to reach the target tile. Player's would often perform what was referred to as a "head fake" by changing directions rapidly just as Pinky was approaching an intersection. By displacing the target tile just as Pinky	
Small model completion	**GPT-2 completion**
was coming to the left, they would get the "correct" direction. If the location is actually farther from what Pac-Man was at, the player's will not respond to the ghost's movement, and the ghost will instead return to the player's current location. This is because Pac-Man does not move by turning. The ghost will have to go to another location of some kind to return where he began, and this is achieved in many ways. It will often simply leave the location, as for example if Pac-Man did jump off a cliff, he would return to the spot where he was before vanishing into space. There is a good number of possibilities these ghost paths may be possible, but the most obvious ones are if the player wants to avoid getting caught up in the game, with a goal as easy as keeping Pac-Man in the game for 1/3 as long as he was. Chasing In Pac-Man's case, this would be similar to how Pac-Man can "catch up" in a fight. A ghost could chase through enemies. This is useful for many reasons, both for Pac-Man and his AI. Unlike most Ghost games, this ghost simply travels in the direction from	was about to cross the intersection, Pac-Man would be able to dodge the ghosts' projectiles and return to the safe location of the safe house. Buster The yellow ghost's AI is specifically designed to try to avoid Pac-Man from leaving the safe house. The player must press the SPACE bar to reveal the safe house's doors, then go into them to find the yellow ghost hiding in the next room. The player must also note that the ghost will not attempt to exit the safe house if it sees Pac-Man already there. In fact, the ghost will not attempt an exit in the first place unless they're already in the safe house. These two behaviors are based on the fact that the player does not always attempt to reach the safe house before Pac-Man (or Blinky). The player must also beware of this glitch if a ghost (or another player) is already in the safe house and attempting to get out as soon as Pac-Man or Blinky is gone. At this point, the ghost will attempt to exit the safe house in a similar manner as it is in "Chase" mode. Pac-Man There are some ghosts that follow Pac-Man. These ghosts are

Table 9. Random unseen contexts (top), and non-cherry-picked completions from both the smallest (left) and largest (right) models. Contexts are 768 tokens, with approximately 256 tokens worth of paragraphs shown. Completions are 256 tokens and fully shown. Top-k random sampling with $k = 40$ was used for generation.

Context (WebText test)	
Corporal Michael P. Goeldin was an unskilled laborer from Ireland when he enlisted in Company A in November 1860. Goldein survived the war. Corporal Patrick O'Neal, also from Ireland, first enlisted in 1854 and served with Company L, 3d U.S. Artillery, in Oregon. He returned to the East Coast and enlisted in the company in 1860. O'Neal served until 1874, when he was named superintendent of the National Cemetery at Willets Point, New York. Corporal Benjamin Browne was a shoemaker from Orange County, New York. In August 1862, he enlisted in the newly formed 124th New York Volunteers, and was one of sixty-one men who transferred into Company A that October. Browne reenlisted in the company in February 1864 while it was camped at Brandy Station. He returned to civilian life after completing his enlistment in 1867. On 10 June, Artificer William Collins was promoted to corporal, probably to fill a combat leadership void for the crossing of the James River. Collins's service record does not reflect the qualities he demonstrated to earn this promotion, but he had obviously overcome some serious problems. Born in Sacketts Harbor, New York, Collins enlisted in the company in December 1853 at the age of twenty-two, and reenlisted in December 1858. Just a month before the war began in April 1861, Collins went "over the hill" and was not caught until three years later. Returned to the company on 22 March 1864, he was tried	
Small model completion	**GPT-2 completion**
on military charges and was discharged from the U.S. Military. He returned to the company, and was assigned to Division of Internal Affairs during the Civil War. As an engineer, Collins also had training in engineering from his native New York. After graduate studies in the Union College of Engineering, he was employed in the United States Army for two years at the same time. There he worked under Captain Daniel Pfeiffer and was promoted to colonel in May 1870, three years after Captain Pfeiffer's enlistment. Collins returned home to England after completing his first year in the First Division to serve the next two years in the Army Corps of Engineers, and by his second year, he began to experience his service duty.¡—endoftext—¿MOSCOW, July 5 (Reuters) - U.K.-based Russian state media and government sources and experts say Moscow has sought to downplay the "military buildup" of the Syrian opposition while maintaining support for President Donald Trump's plan to build a wall along the U.S.-Mexico border, calling it a "military coup". U.S.-led Syria war veteran and Syria envoy Michael Flynn (R), right, gestures with U.S.-based Syrian political analyst Ahmet Koryat, with	in June of that year. Sentenced to be shot for desertion, he remained on active duty. Captain James J. Dall was born in Virginia in 1829 during the Revolution and joined his father in the Army in 1836. He served with the 2d Virginia Infantry until 1845, when he went to the Western Reserve Regiment; and, on his discharge, he reenlisted with Company I, 25th New York Volunteers (which in 1845 numbered about 4,500 men). One of his brothers, James, served with his company in the same capacity in the same brigade. While the regiment remained in Connecticut through 1846, Captain Dall served with Company I at various times before his discharge. A veteran of the Mexican War, he had served two years on the staff of Brigadier General John J. Sullivan at San Antonio. During the Peninsula Campaign of 1863, he commanded a company of twenty-three and was in charge of the battery of fifty-four heavy mortars and the gunboat Pinta of the Battery, a gunboat and several other boats. Captain Dall was on active duty with Company I in the fall of 1865. Two years later, he was promoted to second lieutenant. Captain John A. O'

Table 10. Random unseen contexts (top), and non-cherry-picked completions from both the smallest (left) and largest (right) models. Contexts are 768 tokens, with approximately 256 tokens worth of paragraphs shown. Completions are 256 tokens and fully shown. Top-k random sampling with $k = 40$ was used for generation.

Context (WebText test)

Xmonad, Xresources, and Fonts

Xresources can be an absolute pain (they were for me). So, this a bit of help:

First of all, the way that fonts are generally handled in Xmonad is through XFT . Xft is the X FreeType Interface library and is a library that handles the actual rendering of fonts in Xmonad/urxvt (Xresources specifies a bunch of resources to things that launch under the X window manager. Settings for individual programs are specified in Xresources like so: Urxvt.background:background settins here).

Anyway, fonts are set using this syntax in Xresources :

urxvt*font:xft:your font name goes here:size=the size of the font goes here,xft:fall-back font 1 goes here

So, thats the most basic part of fonts in urxvt specified via Xresources.

Fixing /.ssh/ Permissions in Cygwin

In Cygwin on Windows I found that I could not set the permissions of my /.ssh/ folder to be 0600, as is required for ssh to allow you to use keys. The symptom I had was that I no matter what I did, it always modfied the owner and the group

Small model completion

was set to /.ssh/ , which was the way things were. This could not be fixed completely. I found I could make that much more workable with just typing the username with your mouse then trying to set permissions to be ”0600” for the group, which is what my system is now configured to do.

However, as of Cygwin 7 on Linux I have found that doing so is now ”unnecessary” for my system! I also installed a package called chroot. It does that by allowing you to install/upgrade to your system directory.

So, as of Windows 8 and up, installing that can require chroot, so I tried and tried and found that it wasn’t so bad when installing the package on Linux and then uninstalling it. As of my last write-up on Cygwin, the script just does its job (the sudo commands do the rest), if you try it you will have to log the change.

If that wasn’t confusing enough, this is what I tried and just went back to my /.ssh/ directory and the chroot script. I found I was finally running out of permissions (because i forgot to create a symlink to ssh from the file system and

GPT-2 completion

, making these files read-only. I tried just starting ssh and using a program like passwd which modifies the owner/group to 0644 permissions, but it didn’t work. It was annoying but not nearly as bad as it could be, just annoying to have to fiddle with permissions.

So I decided to do this as the reverse:

For each .ssh folder in your Cygwin directory, set the permission of each file and directory to 2644. Then start ssh with: ssh-add /.ssh/

The first time you connect, your permissions of the /.ssh folder that you just had in Cygwin will be set to 0600. Then every time you connect again, your permissions will change to 0644.

Makefile for Xresources as used in /.xsessionrc

These are made in my .xsessionrc :

.xprofile: .xsession-errors: color=””

For example:

.xprofile: color=blue

As the colors are stored in a separate file, the color in your /.xsession-errors will look like cyan instead of blue.

This is mostly because, as of now, only the color is assigned to your

Table 11. Random unseen contexts (top), and non-cherry-picked completions from both the smallest (left) and largest (right) models. Contexts are 768 tokens, with approximately 256 tokens worth of paragraphs shown. Completions are 256 tokens and fully shown. Top-k random sampling with $k = 40$ was used for generation.

Context (WebText test)		
Aaron loves mint chocolate cake, but he requires that it be paired with mini chocolate chips, so I threw some of those in between the layers. I also had a few Peppermint Jo Jos on hand so I crushed them up and threw some of those in along with some crushed meringue cookies because, why not? It's a total smorgasbord of minty chocolate chippy cookie crunchy goodness. I didn't measure how much of each topping I used, but after I tasted the finished product, I wish I had added more. You can add anything you want- crushed candy canes, peppermint bark, etc. And don't be afraid to use a heavy hand. Texture = good. If you don't have 7-inch cake pans, you can get 3 shorter 8-inch layers out of this		
GPT-2 Completion 1	**GPT-2 Completion 2**	**GPT-2 Completion 3**
recipe. I made the longer version, then used the pan for a layer of buttercream frosting and stuck 1/3 of it into the fridge for a frosting topper. Just make sure you have all of this frosting prepared earlier than you make this cake and you won't have to wait on it as your egg whites sit on the bottom of your mixing bowl. You want the egg whites whipped to stiff peaks? Then put your cake dish over a pot of simmering water and heat it up until it's just barely warm. It doesn't take more than 3-5 minutes; just enough to get it started	recipe. If you do, and you know how to make that cake, then your life will never be the same. I have two very simple rules when it comes to cookie crumbs or chocolate chips. 1) When I buy raw sugar, I do not use it until it's cooled down and has at least half a teaspoon of sugar on it. That's my rule. And 2) If I have a recipe that calls for something and I want to use it in a recipe, I do not substitute it with something else on-line. Like this. I found this really cute card with cute little kittens on	recipe, which will likely yield about the same amount of cake. Chocolate Buttermilk Cake Print Chunks of chocolate cake, coated in buttermilk and topped with crunchy chopped chocolate, drizzled with chocolate sauce. Simple but a nice change from chocolate cake! Author: Momma Made Ingredients For the chocolate buttermilk cake 1 1/2 cup unsalted butter , room temperature 1 1/2 cup granulated sugar 3 large eggs plus 1 egg yolk For the chocolate glaze 1/3 cup cocoa powder 1 3/4 cups powdered sugar 6 ounces semis
GPT-2 Completion 4	**GPT-2 Completion 5**	**GPT-2 Completion 6**
recipe. Meringue Cookies 1 3/4 cup butter, softened 1 cup sugar 1 egg yolk 3 T. heavy cream 3 1/2-4 cups flour pinch of salt Peppermint-Jojo Topping 1 1/4 cups powdered sugar 1 cup chopped pecans 1/2 cup finely chopped mint leaves 1/2 cup chopped fresh mint (about 1/2 sheet) 1 tsp. peppermint extract 1/4 tsp. sea salt pinch of sea	recipe. My 2 failures at Christmas morning were not due to poor frosting skills. That was another success story. The cake was beautifully moist, and well, just a beautiful cake. It's got very little air in it or moisture, but all the chocolate and mint was perfectly blended together. One thing that made me cringe a bit was that all those ingredients- egg yolks, sugar, melted butter, chocolate chips, melted chocolate, and mint- had to be mixed right into the cake batter, which is super time-consuming. A friend suggested that if I put the batter in a separate bowl, I	recipe, which will give you about 6-9 servings of delicious chocolate cake. To prevent your layers from sticking together, I used my 8-inch round tart pan, lined it with parchment paper, then cut up 4 7-inch round cakes from it and sandwiched the layers between them (this also worked well for the 1-layer chocolate cake as well). When you're ready to bake this cake, remove the parchment paper, preheat oven to 350 degrees and bake for about 30 – 40 minutes, until the top is golden and a tester comes out clean. Chocolate Buttermilk Layer Cake with Peppermint Jo

Table 12. Non-cherry-picked completions from GPT-2 generated from the same context (from WebText test). Context is 384 tokens (shown truncated), and generations are 128 tokens. Top-k random sampling with $k = 40$ was used for generation.

Context (human-written): In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

GPT-2: The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.

While examining these bizarre creatures the scientists discovered that the creatures also spoke some fairly regular English. Pérez stated, "We can see, for example, that they have a common 'language,' something like a dialect or dialectic."

Dr. Pérez believes that the unicorns may have originated in Argentina, where the animals were believed to be descendants of a lost race of people who lived there before the arrival of humans in those parts of South America.

While their origins are still unclear, some believe that perhaps the creatures were created when a human and a unicorn met each other in a time before human civilization. According to Pérez, "In South America, such incidents seem to be quite common."

However, Pérez also pointed out that it is likely that the only way of knowing for sure if unicorns are indeed the descendants of a lost alien race is through DNA. "But they seem to be able to communicate in English quite well, which I believe is a sign of evolution, or at least a change in social organization," said the scientist.

Table 13. Conditional generation on an out-of-distribution context by GPT-2. Cherry pick of 10 samples generated with $k = 40$.

<table>
<tr><td colspan="2">Article: Prehistoric man sketched an incredible array of prehistoric beasts on the rough limestone walls of a cave in modern day France 36,000 years ago.
Now, with the help of cutting-edge technology, those works of art in the Chauvet-Pont-d'Arc Cave have been reproduced to create the biggest replica cave in the world.
The manmade cavern named the Caverne du Pont-d'Arc has been built a few miles from the original site in Vallon-Pont-D'arc in Southern France and contains 1,000 painstakingly-reproduced drawings as well as around 450 bones and other features...
Cavemen and women sketched an incredible array of prehistoric beasts on the rough limestone walls of a cave 36,000 years ago and now a replica has been created (pictured)
...</td></tr>
<tr><td>GPT-2: The original site in Vallon-Pont-D'arc in Southern France is a Unesco World Heritage site and is the oldest known and the best preserved cave decorated by man. The replica cave was built a few miles from the original site in Vallon-Pont-D'Arc in Southern France. The cave contains images of 14 different species of animals including woolly rhinoceros, mammoths, and big cats.</td><td>Reference: Cave mimics famous Caverne du Pont-d'Arc in France, the oldest cave decorated by man and the best preserved. The replica contains all 1,000 paintings which include 425 such as a woolly rhinoceros and mammoths. Minute details were copied using 3D modelling and anamorphic techniques, often used to shoot widescreen images. The modern cave also includes replica paw prints of bears, bones and details preserved in the original cave.</td></tr>
<tr><td colspan="2">Article: 'If you can find my submarine, it's yours,' Russian oil billionaire Roman Abramovich once said.
And it seems the oligarch and Chelsea FC owner, whose submarine is just one of the extras that came with his £300million superyacht Eclipse (perfect for getting to shore undetected), is not the only wealthy businessman splashing out on underwater exploration.
Dubbed Earth's real 'final frontier', the oceans are still so little-explored that billionaires are queuing up to buy vessels that give them a glimpse of the dramatic seascapes and incredible wildlife of the world's oceans.
So if you have a spare few million in the bank and want some holiday snaps and Instagram posts that will really trump everyone else, you may want to snap up one of these...
Whale of a time: The OrcaSub takes you 2000 feet beneath the surface in two pressurised Perspex viewing domes for optimum exploration
...</td></tr>
<tr><td>GPT-2: The U-Boat Worx submarine can dive to 984 feet (300 metres) and is capable of diving to the bottom of the ocean. It can be used for diving, snorkelling and even snorkel diving. The submarine is also equipped with a snorkelling platform.</td><td>Reference: Feel like a James Bond villain with the most advanced submerged ocean vehicles. Some convert from a yacht into a submarine, and others detached off for a speedy exploration. Oliver's Travels offers a Mile Low package on its submarine, complete with chef and butler for £175,000 a night.</td></tr>
<tr><td colspan="2">Article: Amina Ali Qassim is sitting with her youngest grandchild on her lap, wiping away tears with her headscarf. Only a few months old, this is the baby girl whose ears she desperately tried to cover the night the aerial bombardment started. She lay awake, she says, in a village mosque on the Yemeni island of Birim, counting explosions as the baby cried.
It could have been worse though. They could have still been in their house when the first missile landed.
"Our neighbor shouted to my husband 'you have to leave, they're coming.' And we just ran. As soon as we left the house, the first missile fell right by it and then a second on it. It burned everything to the ground," Qassim tells us
...</td></tr>
<tr><td>GPT-2: Yemen is in the middle of a civil war. Saudi Arabia is leading the coalition bombing campaign. It's been bombing Yemen for more than two months now.</td><td>Reference: Amina Ali Qassim's family sought shelter in a mosque before fleeing Yemen. Thousands like them are boarding boats to sail to Djibouti. Saudi Arabia has been pounding Yemen in a bid to defeat Houthi rebels.</td></tr>
</table>

Table 14. Summaries generated by GPT-2 and reference summaries on the test set of the CNN and Daily Mail dataset.

English reference This re-release, titled The Next Day Extra, was presented in the form of three disks: the original album, unpublished studio sessions and remixes, plus a DVD containing the four clips that have already been unveiled.	**GPT-2 French translation** Les nouvelles re-releases, tout en premier disc, nécessaire de l'album, un studio session et remixes, plus une DVD de l'écran de quelques clips qui ont été déjà échappés.
French reference Un homme expliquait que le fonctionnement de la hernia fonctionnelle qu'il avait reconnaît avant de faire, le fonctionnement de la hernia fonctionnelle que j'ai réussi, j'ai réussi.	**GPT-2 English translation** This release, originally released on three discs: the original album, sessions studio inédites and remixes, a DVD containing the four clips that were released on the previous day.
English reference One man explained that the free hernia surgery he'd received will allow him to work again.	**GPT-2 French translation** Un homme expliquait que le fonctionnement de la hernia fonctionnelle qu'il avait reconnaît avant de faire, le fonctionnement de la hernia fonctionnelle que j'ai réussi, j'ai réussi.
French reference Un homme a expliqué que l'opération gratuite qu'il avait subie pour soigner une hernie lui permettrait de travailler à nouveau.	**GPT-2 English translation** A man told me that the operation gratuity he had been promised would not allow him to travel.
English reference Speaking to an open government conference in London via video link, Mr Kerry said: "There is no question that the President and I and others in government have actually learned of some things that had been happening on an automatic pilot because the ability has been there, going back to World War Two and to the very difficult years of the Cold War, and then, of course, 9/11."	**GPT-2 French translation** Le premier conference de l'édition internationale en ligne sur la vide, le ministre de l'Édition internationale a été vide, lui dit : "Il n'y a qu'une question que le président et moi et tous les autres personnes d'une administration qui a été évident par la fois, par la part de la guerre, par la part de la guer
French reference S'exprimant lors d'une conférence intergouvernementale à Londres par liaison vidéo, M. Kerry a déclaré: "Il est indéniable que le Président, moi-même et d'autres membres du gouvernement avons pris connaissance de certaines choses en mode pilote automatique parce que nous en avions la possibilité, dès la Seconde guerre mondiale et jusqu'aux années difficiles de la Guerre froide, puis bien sûr le 11 septembre."	**GPT-2 English translation** In a conférence between the United States and London, Secretary of State John Kerry said: "It is indeniable that the President, myself and others of the government have been aware of certain certain choices that have been made in the past in order to be able to do certain things in a more automated way."

Table 15. English to French and French to English translations generated by GPT-2.

Context (passage and previous question/answer pairs)

The 2008 Summer Olympics torch relay was run from March 24 until August 8, 2008, prior to the 2008 Summer Olympics, with the theme of "one world, one dream". Plans for the relay were announced on April 26, 2007, in Beijing, China. The relay, also called by the organizers as the "Journey of Harmony", lasted 129 days and carried the torch 137,000 km (85,000 mi) – the longest distance of any Olympic torch relay since the tradition was started ahead of the 1936 Summer Olympics.

After being lit at the birthplace of the Olympic Games in Olympia, Greece on March 24, the torch traveled to the Panathinaiko Stadium in Athens, and then to Beijing, arriving on March 31. From Beijing, the torch was following a route passing through six continents. The torch has visited cities along the Silk Road, symbolizing ancient links between China and the rest of the world. The relay also included an ascent with the flame to the top of Mount Everest on the border of Nepal and Tibet, China from the Chinese side, which was closed specially for the event.

Q: What was the theme
A: "one world, one dream".

Q: What was the length of the race?
A: 137,000 km

Q: Was it larger than previous ones?
A: No

Q: Where did the race begin?
A: Olympia, Greece

Q: Is there anything notable about that place?
A: birthplace of Olympic Games

Q: Where did they go after?
A: Athens

Q: How many days was the race?
A: seven

Q: Did they visit any notable landmarks?
A: Panathinaiko Stadium

Q: And did they climb any mountains?
A:

Model answer: Everest
Turker answers: unknown, yes, Yes, yes

Table 16. Selected CoQA completion.

Context (passage and previous question/answer pairs)

Tom goes everywhere with Catherine Green, a 54-year-old secretary. He moves around her office at work and goes shopping with her. "Most people don't seem to mind Tom," says Catherine, who thinks he is wonderful. "He's my fourth child," she says. She may think of him and treat him that way as her son. He moves around buying his food, paying his health bills and his taxes, but in fact Tom is a dog.

Catherine and Tom live in Sweden, a country where everyone is expected to lead an orderly life according to rules laid down by the government, which also provides a high level of care for its people. This level of care costs money.

People in Sweden pay taxes on everything, so aren't surprised to find that owning a dog means more taxes. Some people are paying as much as 500 Swedish kronor in taxes a year for the right to keep their dog, which is spent by the government on dog hospitals and sometimes medical treatment for a dog that falls ill. However, most such treatment is expensive, so owners often decide to offer health and even life _ for their dog.

In Sweden dog owners must pay for any damage their dog does. A Swedish Kennel Club official explains what this means: if your dog runs out on the road and gets hit by a passing car, you, as the owner, have to pay for any damage done to the car, even if your dog has been killed in the accident.

Q: How old is Catherine?
A: 54

Q: where does she live?
A:

Model answer: Stockholm
Turker answers: Sweden, Sweden, in Sweden, Sweden

Table 17. Selected CoQA completion.