

Language Models are Unsupervised Multitask Learners

Alec Radford^{*1} Jeffrey Wu^{*1} Rewon Child¹ David Luan¹ 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

语言模型是无监督的多任务学习者

亚历克·拉德福^{d * 1} Jeffrey Wu^{* 1} Rewon Child¹ David Luan¹ Dario Amodei^{** 1} Ilya Sutskever^{** 1}

摘要

自然语言处理任务（如问答、机器翻译、阅读理解和摘要生成）通常通过在特定任务数据集上进行监督学习来实现。我们证明，当语言模型在名为WebText的数百万网页新数据集上训练时，即使没有显式监督，也能开始学习这些任务。当语言模型在文档加问题条件下生成答案时，其在CoQA数据集上达到55%的F1分数——在未使用127,000+ 训练样本的情况下，该表现可匹敌或超越四种基线系统中的三种。语言模型的容量是零样本任务迁移成功的关键，其容量提升能以对数线性方式改善跨任务性能。我们最大的模型GPT-2采用15亿参数的Transformer架构，在零样本设置下于8个测试语言建模数据集中有7个取得最先进结果，但仍对WebText数据存在欠拟合。模型生成的文本样本体现了这些改进，包含连贯的段落结构。这些发现为构建语言处理系统指明了前景广阔的道路——该系统能从自然生成的示范中学习执行任务。

我们希望朝着更通用的系统发展，使其能够执行多种任务——最终无需为每个任务手动创建和标注训练数据集。

构建机器学习系统的主流方法是：收集展示目标任务正确行为的训练样本数据集，训练系统模仿这些行为，然后在独立同分布（IID）的保留样本上测试其性能。这种方法在窄领域专家系统中成效显著。然而，在面对多样化输入时，描述模型（Lake et al., 2017）、阅读理解系统（Jia & Liang, 2017）和图像分类器（Alcorn et al., 2018）常表现出不稳定行为，这凸显了该方法的局限性。

我们推测，当前系统普遍采用单领域数据集进行单任务训练，是导致其泛化能力不足的主要原因。要基于现有架构构建稳健系统，可能需要在广泛领域和任务上进行训练并评估性能。近期已出现若干基准测试框架（如GLUE（Wang等，2018）和decaNLP（McCann等，2018））来开展相关研究。

多任务学习（Caruana, 1997）是提升整体性能的有效框架，但在NLP领域仍处于萌芽阶段。近期研究显示性能提升幅度有限（Yogatama et al., 2019），迄今最具雄心的两项工作分别在10个和17个（数据集，目标）组合上进行训练（McCann et al., 2018）（Bowman et al., 2018）。从元学习视角看，每个（数据集，目标）对都是从数据集和目标分布中抽取的单个训练样本。当前机器学习系统需要数百至数千个样本才能诱导出具有良好泛化能力的函数。这表明多任务训练可能同样需要大量有效训练对才能实现其潜力。若仅依靠现有技术进行蛮力突破，持续扩大数据集创建规模和目标设计范围将面临巨大困难。这促使我们探索更多多任务学习的实现方案。

当前语言任务领域表现最优的系统

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

^{*}、^{**}同等贡献 ¹OpenAI, 旧金山, 加利福尼亚州, 美国。通讯作者: Alec Radford<alec@openai.com>。

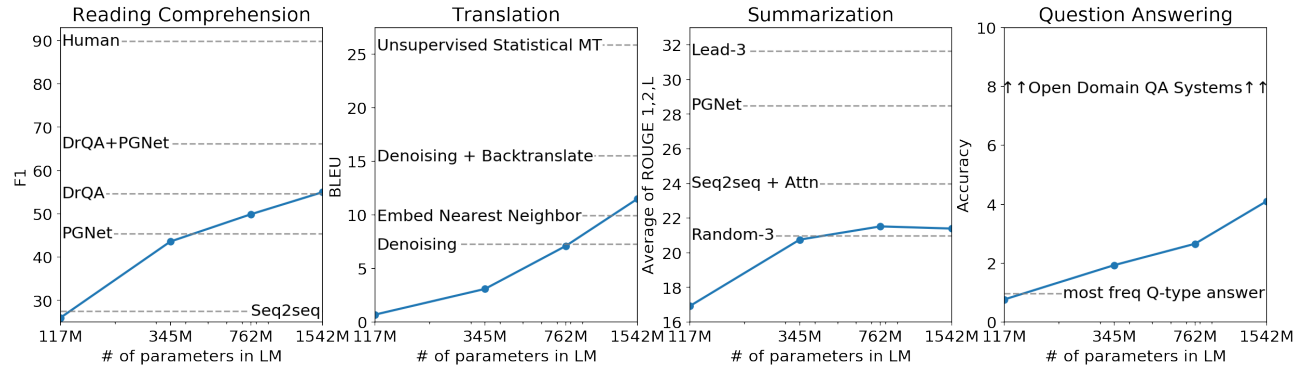


Figure 1. Zero-shot task performance of WebText LMs as a function of model size on many NLP tasks. Reading Comprehension results are on CoQA (Reddy et al., 2018), translation on WMT-14 Fr-En (Artetxe et al., 2017), summarization on CNN and Daily Mail (See et al., 2017), and Question Answering on Natural 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, \dots, x_n) each composed of variable length sequences of symbols (s_1, s_2, \dots, 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):

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

This approach allows for tractable sampling from and estimation of $p(x)$ as well as any conditionals of the form $p(s_{n-k}, \dots, s_n | s_1, \dots, s_{n-k-1})$. 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,

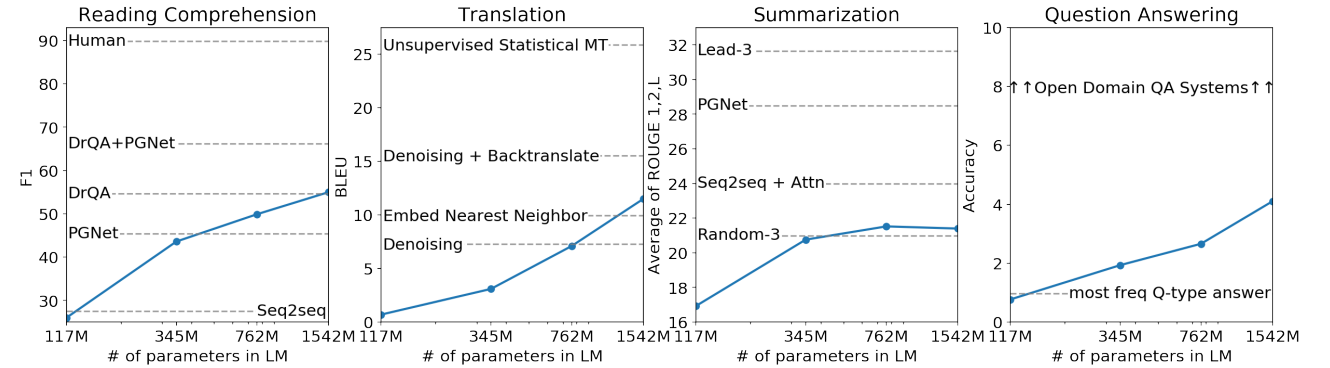


图1. WebText语言模型在多NLP任务上的零样本任务表现随模型规模变化曲线。阅读理解结果基于CoQA (Reddy等人, 2018)，翻译任务采用WMT-14法语-英语数据集 (Artetxe等人, 2017)，摘要生成使用CNN和每日邮报数据集 (See等人, 2017)，问答任务则基于Natural Questions数据集 (Kwiatkowski等人, 2019)。第3节包含对每项结果的详细描述。

采用预训练与监督式微调相结合的方法。这种方法历史悠久，且呈现出向更灵活迁移形式发展的趋势。最初通过学习词向量作为任务特定架构的输入 (Mikolov et al., 2013) (Collobert et al., 2011)，随后转移循环网络的上下文表示 (Dai & Le, 2015) (Peters et al., 2018)，而近期研究表明特定任务架构已非必需，转移多个自注意力模块即可满足需求 (Radford et al., 2018) (Devlin et al., 2018)。

这些方法仍需依赖监督训练才能执行任务。当监督数据稀缺或缺失时，另一类研究表明语言模型在特定任务中展现出潜力，例如常识推理 (Schwartz et al., 2017) 和情感分析 (Radford et al., 2017)。

本文将这两条研究路径相融合，延续了更通用迁移方法的发展趋势。我们证明语言模型能在零样本环境下执行下游任务——无需任何参数或架构调整。通过凸显语言模型在零样本环境下执行广泛任务的能力，我们展示了该方法的潜力。根据任务类型不同，我们取得了令人鼓舞的、具有竞争力的、以及处于技术前沿的成果。

2. 研究方法

我们方法的核心是语言建模。语言建模通常被定义为从一组示例 (x_1, x_2, \dots, x_n) 中进行无监督分布估计，每个示例由可变长度的符号序列 (s_1, s_2, \dots, s_n) 组成。由于语言具有天然的序列结构，常见做法是分解联合概率：

符号作为条件概率的乘积 (Jelinek & Mercer, 1980) (Bengio et al., 2003)：

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

该方法可实现 $p(x)$ 的可处理采样与估计，以及任何形式为 $p(s_{n-k}, \dots, s_n | s_1, \dots, s_{n-k-1})$ 的条件概率计算。近年来，能够计算这些条件概率的模型在表达能力方面取得了显著进步，例如Transformer (Vaswani等人, 2017) 等自注意力架构。

单任务学习可通过概率框架表述为条件分布 $p(output | input)$ 的估计。由于通用系统需处理多种任务（即使面对相同输入），其建模应同时考虑输入与目标任务，即建模 $p(output | input, task)$ 。这一思想已在多任务和元学习场景中得到多种形式化表述。任务条件化常在架构层面实现，如Kaiser等人 (2017) 提出的任务专用编码器与解码器；或在算法层面实现，如MAML (Finn等人, 2017) 的内循环与外循环优化框架。但正如McCann等人 (2018) 所示例，语言提供了一种灵活的方式，可将任务、输入和输出统统表述为符号序列。例如，翻译训练样本可写为序列(翻译成法语, 英文文本, 法语文本)。同理，阅读理解训练样本可表述为(回答问题, 文档, 问题, 答案)。McCann等人 (2018) 证明了通过MQAN模型可实现单模型训练，

Language Models are Unsupervised Multitask Learners
<p>to infer and perform many different tasks on examples with this type of format.</p> <p>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.</p> <p>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.</p>
<p>”I’m not the cleverest man in the world, but like they say in French: Je ne suis pas un imbécile [I’m not a fool].</p> <p>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.”</p> <p>“I hate the word ‘perfume,”’ Burr says. ‘It’s somewhat better in French: ‘parfum.’</p> <p>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?”.</p> <p>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?</p> <p>“Brevet Sans Garantie Du Gouvernement”, translated to English: “Patented without government warranty”.</p>

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 Newspaper¹ 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>

语言模型是无监督的多任务学习者
<p>通过此类格式示例推断并执行多种不同任务。</p> <p>语言模型在理论上能够学习McCann等人（2018）提出的任务，而无需明确监督需预测的输出符号。由于监督目标与无监督目标相同，但仅在序列子集上进行评估，因此无监督目标的全局最小值即为监督目标的全局最小值。 在此简化的实验设置中，(Sutskever et al., 2015)文中关于密度估计作为原理性训练目标的担忧得以规避。问题转而聚焦于实践中能否将无监督目标优化至收敛状态。初步实验证实，足够庞大的语言模型能在该简化设置中实现多任务学习，但学习速度远低于显式监督方法。</p> <p>尽管从上述规范化设置到"自然语言环境"的复杂性存在巨大鸿沟，Weston（2016）在对话领域论证了开发直接学习自然语言系统的必要性，并通过"利用教师输出前向预测学习问答任务"的验证性研究（无需奖励信号）证明了该概念的可行性。虽然对话方法极具吸引力，但我们担忧其限制性过强。 互联网蕴藏着海量无需交互即可被被动获取的信息。 我们推测：当语言模型具备足够容量时，将开始学习推断并执行自然语言序列中展示的任务，以提升预测精度——无论这些序列的获取方式如何。若语言模型能实现此能力，实质上即在执行无监督多任务学习。我们通过分析语言模型在零样本环境下处理各类任务的表现，验证这一假设是否成立。</p>
<p>"我并非世上最聪明的人，但正如法语所言：J'en suis pas un imbécile [我并非愚人]。"</p> <p>在8月16日现已删除的帖文中，乔利埃特选区保守党候选人索海尔·艾德用法语写道："Mentez mentez, il en restera toujours quelque chose", 意为"谎言谎言，总会留下些什么"。</p> <p>"我讨厌'perfume'这个词，"伯尔说，"法语'parfum'要好些。"</p> <p>若在29:55处仔细聆听，可听到两名男子用法语对话："- Comment on fait pour aller de l'autre cot'e? - Quel autre cot'e?", 意为"-怎么去对岸? -哪边? "。</p> <p>若觉得上述说明稍显牵强，不妨参考这个法语问题：As-tu aller au cin'ema?（你去看电影了吗？），其字面翻译为"你-去-电影/剧院"。</p> <p>“Brevet Sans Garantie Du Gouvernement”，译为英文：“Patented without government warranty”。</p>

表1. 摘自WebText训练集的英语→法语与法语→英语自然语境翻译示例。

Common Crawl数据集。Trinh & Le（2018）取得的最佳结果是通过Common Crawl的小型子样本实现的，该子样本仅包含与目标数据集（Winograd Schema Challenge）最相似的文档。虽然这是提升特定任务性能的实用方法，但我们希望避免预先对待执行任务做出假设。

为此我们创建了新的网页抓取方案，重点提升文档质量。具体做法是仅抓取经过人工筛选/过滤的网页。手动筛选全部网页数据成本极高，因此我们首先抓取了社交媒体平台Reddit上所有获得至少3点声望值的外链。这可视为一种启发式指标，反映其他用户是否认为该链接有趣、有教育意义或纯粹好笑。

最终生成的WebText数据集包含这4500万链接的文本子集。为从HTML响应中提取文本，我们结合使用了Dragnet（Peters & Lecocq, 2013）和Newspaper¹内容提取器。 本文所有结果均基于WebText的初始版本，该版本未包含2017年12月后创建的链接，经去重及启发式清理后包含略超800万份文档，文本总量达40GB。我们从WebText中移除了所有维基百科文档，因其作为其他数据集的常见来源，可能因过度使用导致分析复杂化。

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 lower-casing, 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 sub-optimal 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

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.

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.

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

将训练数据与测试评估任务进行混合处理。

2.2. 输入表示

通用语言模型（LM）应能计算任意字符串的概率（并能生成该字符串）。当前大规模语言模型包含预处理步骤（如小写转换、分词及词表外令牌处理），这些步骤限制了可建模字符串的空间。将Unicode字符串作为UTF-8字节序列处理可优雅满足此要求，[Gillick等人（2015）](#)等研究即为明证。[\(2015\)](#)等研究中展示的那样，优雅地满足了这一要求。然而在大型数据集（如十亿词基准测试Al-Rfou等人，2018）上，当前字节级语言模型仍无法与词级语言模型抗衡。我们在尝试用WebText数据集训练标准字节级语言模型时，也观察到了类似的性能差距。

字节对编码（BPE）（[Sennrich等人，2015](#)）是字符级与词级语言建模之间的实用折中方案，能有效地在高频符号序列的词级输入与低频符号序列的字符级输入之间进行插值。尽管名称如此，参考BPE实现通常操作于Unicode码点而非字节序列。这些实现需包含完整的Unicode符号空间才能建模所有Unicode字符串，这将导致基础词汇量在添加多符号标记前就超过13万。相较于BPE常用的32,000至64,000词汇量，这规模过于庞大。反观字节级BPE仅需256词的基础词汇表。但直接将BPE应用于字节序列会导致次优合并结果，因其采用贪婪频率启发式构建词汇表。我们观察到BPE会包含大量常见词的多重变体（如`dog`、`dog!`、`dog?`），导致有限词汇槽位与模型容量分配效率低下。为避免此问题，我们禁止BPE对任何字节序列进行跨字符类别合并，仅对空格添加例外处理——这在显著提升压缩效率的同时，仅产生极小程度的词汇碎片化（即词汇被拆分至多个词汇表条目）。

这种输入表示方式使我们能够结合词级语言模型的经验优势与字节级方法的普适性。由于我们的方法可为任意Unicode字符串分配概率，因此无论预处理方式、分词规则或词汇量大小如何，我们都能在任意数据集上评估语言模型。

2.3. 模型

我们采用基于Transformer（[Vaswani等，2017](#)）的架构构建语言模型。该模型基本遵循OpenAI GPT模型（[Radford等，2018](#)）的细节设计，并通过

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

表2. 四种模型规模的架构超参数设置。

少量修改。层归一化（[Ba et al., 2016](#)）被移至每个子模块输入端，类似于预激活残差网络（[He et al., 2016](#)），并在最终自注意力模块后新增一层归一化。采用改进的初始化方案，该方案考虑了残差路径随模型深度的累积效应。初始化时，残差层权重按 $1/\sqrt{N}$ 倍缩放（其中 N 为残差层数量）。词汇表扩展至50,257个词。同时将上下文规模从512个词元增至1024个词元，并采用更大的512批量大小。

3. 实验部分

我们训练并评估了四种参数规模近似呈对数均匀分布的语言模型，其架构详见表2。最小模型等效于原始GPT，次小模型等效于BERT（[Devlin等人，2018](#)）的最大模型。我们称之为GPT-2的最大模型参数数量比GPT高出一个数量级。每个模型的学习率均通过手动调优，以在WebText数据集5%保留样本上获得最佳困惑度。所有模型对WebText数据集仍存在欠拟合现象，且随着训练时间增加，保留样本困惑度目前仍在持续提升。

3.1. 语言建模

作为零样本任务迁移的初步探索，我们关注WebText语言模型在其核心训练任务——语言建模——上的零样本领域迁移表现。由于本模型在字节层面运行且无需损耗性预处理或分词操作，可适配任何语言模型基准测试。语言模型数据集的评估指标通常采用缩放或指数化的经典预测单元（字符/字节/词）平均负对数概率。我们通过计算WebText语言模型对数据集的对数概率，再除以经典预测单元数量来实现同等指标。对于多数数据集而言，WebText语言模型将面临显著的分布外测试场景：需预测高度标准化文本、标记化异常（如断开标点与缩写）、打乱的句子，甚至字符串

Language Models are Unsupervised Multitask Learners										
	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

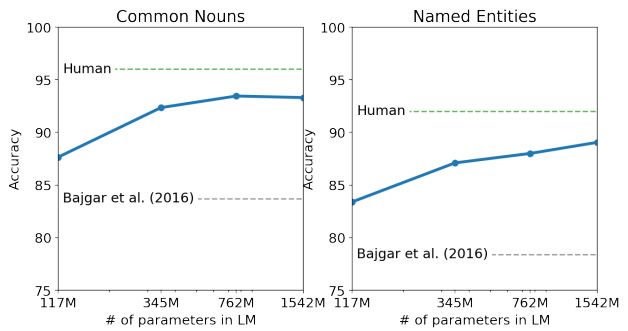


Figure 2. Performance on the Children’s Book Test as a function of model capacity. Human performance are from Bajgar et al. (2016), instead of the much lower estimates from the original paper.

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

语言模型是无监督的多任务学习者										
	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

表3. 多数据集零样本测试结果。所有结果均未进行训练或微调。PTB与WikiText-2结果引自(Gong et al., 2018)。CBT结果引自(Bajgar et al., 2016)。LAMBADA准确率结果引自(Hoang et al., 2018)，LAMBADA困惑度结果引自(Grave et al., 2016)。其余结果均引自(Dai et al., 2019)。

<UNK> 这种情况在WebText数据集中极为罕见——400亿字节数据中仅出现26次。我们在表3中报告了使用可逆去分词器的主要结果，该技术尽可能消除了分词/预处理过程中的伪影。由于这些去分词器具有可逆性，我们仍可计算数据集的对数概率，其本质可视为一种简易的领域适应方案。实验表明，采用这些去分词器后，GPT-2的困惑度提升幅度达2.5至5。

WebText语言模型在跨领域和数据集迁移中表现优异，在零样本设置下于8个数据集集中的7个实现了突破性进展。在 Penn Treebank 和 WikiText-2 等仅含 100-200 万训练令牌的小型数据集上，模型表现显著提升。针对长时序依赖性设计的数据集（如LAMBADA（Paperno等人，2016）和儿童读物测试（Hill等人，2015））同样呈现大幅改进。在十亿词基准（Chelba et al., 2013）上，本模型表现仍显著逊于前人成果。这可能源于该数据集规模最大且预处理破坏性最强——其句子级打乱操作彻底消除了所有长程结构。

3.2. 儿童读物测试

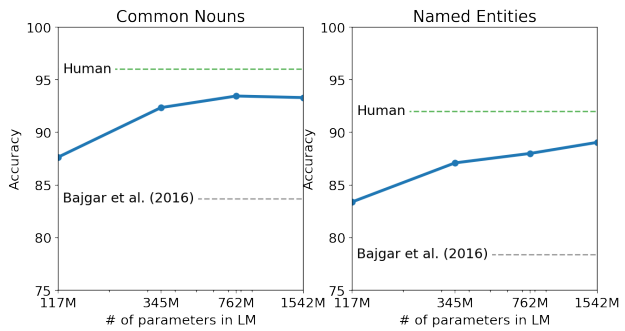


图2. 儿童读物测试中模型容量与性能的关系曲线。人类表现数据源自Bajgar等人（2016）的研究，而非原始论文中更低的预估值。

儿童图书测试（CBT）（Hill et al., 2015）旨在评估语言模型在不同词类（命名实体、名词、动词、介词）上的表现。该测试不采用困惑度作为评估指标，而是通过自动构建的填空题评估准确率——任务要求从10个备选项中预测缺失词的正确答案。遵循原始论文提出的语言模型方法，我们根据语言模型计算每个选项及其对应句子的条件概率，并预测概率最高的选项。如图2所示，随着模型规模增大，性能稳步提升，并基本弥合了与人类表现的差距。数据重叠分析表明，CBT测试集中的《丛林之书》（作者：鲁德亚德·吉卜林）已收录于WebText数据库，因此我们采用无显著重叠的验证集进行结果报告。GPT-2在普通名词识别上达到93.3%的新纪录，命名实体识别率达89.1%。通过去标记化处理，我们消除了CBT数据中PTB风格的标记化伪影。

3.3. LAMBADA

LAMBADA数据集（Paperno等人，2016）用于测试系统建模文本长程依赖关系的能力。该任务要求预测句子的最后一个词，而人类需要至少50个上下文词才能成功预测。GPT-2将该任务的困惑度从99.8（Grave等人，2016）提升至8.6，并将语言模型在此测试中的准确率从19%（Dehghani等人，2018）提高至52.66%。对GPT-2错误的分析表明，多数预测虽能延续句子逻辑，却不符合句末词要求。这说明该模型未采用"句末词必须是上下文词汇"的有效约束。通过添加停用词过滤器作为近似解决方案，准确率进一步提升至63.24%，使该任务的整体技术水平提高4%。先前最先进水平（Hoang等人，2018）采用不同限制预测设置，将模型输出限定在上下文中出现的词汇范围内。对于GPT-2而言，此类限制反而会产生负面影响。

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

3.4. Winograd Schema Challenge

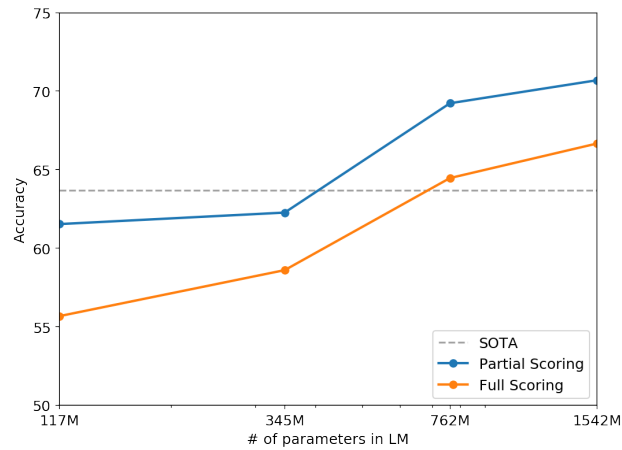


Figure 3. Performance on the Winograd Schema Challenge as a function of model capacity.

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.,

	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)

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*.

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

由于19%的答案与上下文无关，我们采用未经预处理的数据集版本。

3.4. 温诺格拉德模式挑战

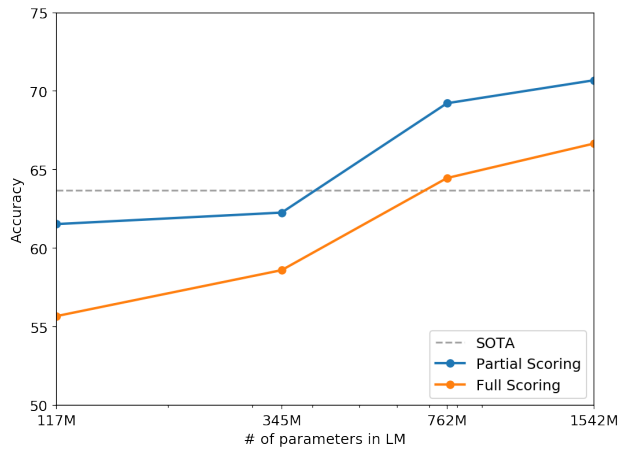


图3. 模型容量对Winograd模式挑战任务表现的影响关系。

Winograd模式挑战（Levesque等人，2012）旨在通过评估系统解决文本歧义的能力，衡量其进行常识推理的水平。近期Trinh与Le（2018）利用语言模型在该挑战中取得重大突破，通过更高概率预测歧义的解决方式。我们遵循其问题框架，并通过图3展示了采用全量与部分计分技术时模型的表现。GPT-2将准确率提升7%至70.70%，超越现有最佳水平。由于该数据集仅含273个样本，建议参阅Trichelair等人（2018）以全面理解该结果。

3.5. 阅读理解

对话问答数据集（CoQA）由Reddy等人（2018）创建，包含来自7个不同领域的文档，并配有提问者与回答者围绕该文档展开的自然语言对话。CoQA既测试阅读理解能力，也检验模型回答依赖对话历史的问题（如“为什么？”）的能力。

基于文档、对话历史及最终令牌A的GPT-2贪婪解码模型，在开发集上达到55%的F1值。该表现与4个基线系统中的3个持平或超越，且无需使用这些基线系统训练时所依赖的127,000+ 人工收集问答对。监督式SOTA系统——基于BERT的模型（Devlin等人，

	R-1	R-2	R-L	R-AVG
自下而上求和	41.22	18.68	38.34	32.75
Lede-3	40.38	17.66	36.62	31.55
Seq2Seq +注意力模型	31.33	11.81	28.83	23.99
GPT-2 核心要点:	29.34	8.27	26.58	21.40
Random-3	28.78	8.63	25.52	20.98
GPT-2 无提示	21.58	4.03	19.47	15.03

表4. 基于CNN和Daily Mail数据集的ROUGE F1指标测算的摘要性能。自下而上摘要（Bottom-Up Sum）采用(Gehrmann et al., 2018)提出的SOTA模型。

2018年）的F1值已接近人类水平（89分）。尽管GPT-2在无监督训练系统中表现令人振奋，但对其答案和错误的分析表明，该模型常采用简单检索式启发策略——例如对“谁”的问题直接引用文档中的姓名作答。

3.6. 文本摘要

我们通过CNN和每日邮报数据集（Nallapati等人，2016）测试GPT-2的摘要生成能力。为诱导摘要行为，我们在文章末尾添加文本“TL;DR:”，并采用Top- k 随机采样（Fan et al., 2018）生成100个令牌，其中 $k = 2$ 参数可减少重复内容，相较贪婪解码更倾向于生成抽象化摘要。我们从这100个令牌中选取前3个生成的句子作为摘要。虽然从定性上来看，这些生成内容类似于摘要（如表14所示），但它们往往聚焦于文章的近期内容，或混淆具体细节，例如事故中涉及多少辆车，或者标志是在帽子上还是衬衫上。在常用的ROUGE 1、2、L指标上，生成的摘要仅勉强接近经典神经网络基线的表现，且仅略优于从文章中随机选取3个句子的效果。当移除任务提示时，GPT-2在综合指标上的表现下降了6.4分，这证明了自然语言模型能够通过任务提示触发特定行为。

3.7. 翻译

我们测试GPT-2是否已开始掌握跨语言翻译能力。为引导模型理解任务目标，我们通过英语句子 = 法语句子形式的例对上下文对语言模型进行条件训练，最终在英语句子 = 的提示下，采用贪婪解码从模型中采样，并将首句生成结果作为译文。在WMT-14英法测试集上，GPT-2获得5分BLEU得分，略逊于先前无监督词翻译研究中采用双语词典进行逐词替换的水平。

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 detector² 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).

³Alec, 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

²<https://github.com/CLD2Owners/cld2>

问题	生成的答案	正确	概率
《物种起源》的作者是谁？	查尔斯·达尔文	✓	83.4%
Ubuntu项目的创始人是谁？	马克·沙特尔沃斯	✓	82.0%
绿湾包装工队的四分卫是谁？	亚伦·罗杰斯	✓	81.1%
熊猫是哪个国家的国兽？	中国	✓	76.8%
谁提出了相对论？	阿尔伯特·爱因斯坦	✓	76.4%
第一部《星球大战》电影何时上映？	1977	✓	71.4%
瑞典最常见的血型是什么？	A	✗	70.6%
谁被视为精神分析学的创始人？	西格蒙德·弗洛伊德	✓	69.3%
1969年，谁在月球上迈出了第一步？	尼尔·阿姆斯特朗	✓	66.8%
英国最大的连锁超市是哪家？	乐购	✓	65.3%
"shalom"在英语中的含义是什么？	和平	✓	64.0%
《孙子兵法》的作者是谁？	孙子	✓	59.6%
美国面积最大的州是哪个？	加利福尼亚州	✗	59.2%
绿藻属于哪种繁殖方式的典型代表？	孤雌生殖	✗	56.5%
维克拉姆历在哪个国家是官方历法？	印度	✓	55.6%
《独立宣言》的主要撰写者是谁？	托马斯·杰斐逊	✓	53.3%
美国哪个州构成蒙大拿州的西部边界？	蒙大拿州	✗	52.3%
《权力的游戏》中谁饰演了戴佛斯爵士？	彼得·丁拉基	✗	52.1%
谁任命联邦储备系统的主席？	珍妮特·耶伦	✗	51.5%
描述将一个细胞核分裂为两个遗传相同的细胞核的过程？	有丝分裂	✓	50.7%
NBA历史上获得最多MVP奖项的球员是谁？	迈克尔·乔丹	✗	50.2%
与罗马城相关的河流是？	台伯河	✓	48.6%
首位遭弹劾的总统是谁？	安德鲁·约翰逊	✓	48.3%
2017年美国国土安全部部长是谁？	约翰·凯利	✓	47.0%
欧盟的通用货币名称是什么？	欧元	✓	46.8%
星球大战中皇帝的名字是什么？	帕尔帕廷	✓	46.5%
在射击场射击是否需要持枪许可证？	不	✓	46.4%
1859年，谁将进化论作为生物发展的基础提出？	查尔斯·达尔文	✓	45.7%
俄罗斯核电站爆炸事件？	切尔诺贝利	✓	45.7%
在原版《终结者》中，谁饰演了约翰·康纳？	阿诺德·施瓦辛格	✗	45.2%

表5. GPT-2在Natural Questions开发集上生成的30个最自信答案，按GPT-2评估的概率排序。根据第4节所述流程，这些问题均未出现在WebText数据集中。

(Conneau et al., 2017b)。在WMT-14法英测试集上，GPT-2凭借其强大的英语语言模型实现了显著提升，获得11.5 BLEU分值。该成绩超越了(Artetxe et al., 2017)和(Lample et al., 2017)提出的若干无监督机器翻译基线模型，但仍远低于当前最佳无监督机器翻译方案(Artetxe et al., 2019)的33.5 BLEU得分。该任务的性能表现令我们意外，因为我们在WebText数据集的过滤步骤中特意剔除了非英文网页。为验证此结论，我们对WebText运行了字节级语言检测器²，结果仅检测到10MB法语数据——这比先前无监督机器翻译研究中常见的单语种法语语料库小约500倍。

3.8. 问答系统

检验语言模型信息承载能力的一种可行方法是评估其对事实类问题生成正确答案的频率。此前在神经系统中展示的此类行为（如所有信息均存储于参数中）——例如神经对话模型（Vinyals & Le, 2015）——因缺乏高质量评估数据集而仅能呈现定性结果。近期推出的Natural Questions数据集（Kwiatkowski et al.,

²<https://github.com/CLD2Owners/cld2>

2019)是测试此类问题更具量化价值的资源。与翻译类似，语言模型的上下文通过示例问答对进行预训练，这有助于模型推断数据集的简答风格。当采用阅读理解数据集（如SQUAD）常用的精确匹配指标评估时，GPT-2正确回答了4.1%的问题。³作为对照，最小模型未能超越极其简单的基线准确率（1.0%）——该基线仅为每种问题类型（谁、什么、何处等）返回最常见答案。GPT-2正确回答的问题数量是其基准模型的5.3倍，表明模型容量是当前神经系统在此类任务中表现欠佳的主要因素。GPT-2对生成答案的置信度分配校准良好，在置信度最高的1%问题上准确率达63.1%。表 5 展示了 GPT-2 在开发集问题上生成的 30 个最自信答案。尽管如此，GPT-2的表现仍远逊于融合信息检索与抽取式文档问答的开放域问答系统（Alberti等, 2019），后者准确率通常在30%至50%区间。

³艾力克曾自认擅长随机知识问答，但在与GPT-2相同测试环境下，其随机抽取的100道题目仅答对17题。实际他仅答对14题，但另外3题本应答对。

Language Models are Unsupervised Multitask Learners						
	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

⁴A 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.

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

语言模型是无监督的多任务学习者						
	PTB	WikiText-2	enwik8	text8	Wikitext-103	1BW
数据集训练	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%

表6. 测试集与训练集重叠的8-gram百分比。

4. 泛化能力与死记硬背

计算机视觉领域的最新研究表明，常见图像数据集中存在大量近似重复图像。例如CIFAR-10训练集与测试集图像存在3.3%的重叠（Barz & Denzler, 2019）。这导致机器学习系统的泛化性能被高估。随着数据集规模扩大，该问题发生概率将显著增加，表明WebText可能存在类似现象。因此分析测试数据在训练数据中的出现比例至关重要。

为研究该课题，我们创建了包含WebText训练集令牌8-gram的布隆过滤器。为提高召回率，字符串经规范化处理，仅保留小写字母数字单词，并以单空格作为分隔符。布隆过滤器的构建确保假阳性率上限为 $\frac{1}{10^{80}}$ 。通过生成100万条字符串验证其低假阳性率，结果显示过滤器未检出任何异常。

这些布隆过滤器使我们能够根据给定数据集，计算其中8-gram在WebText训练集中出现的百分比。表6展示了常见语言模型基准测试集的重叠分析。常见语言模型数据集的测试集与WebText训练集重叠率介于1%-6%之间，平均重叠率为3.2%。值得注意的是，许多数据集与其自身训练集的重叠率更高，平均达5.9%。

我们的方法优化了召回率，虽然人工检查重叠部分发现存在大量常见短语，但许多较长的匹配源于数据重复。这并非WebText独有问题。例如我们发现WikiText-103测试集中存在一篇与训练集重复的文章。由于测试集仅含 60 篇文章，重叠率至少达 1.6%。⁴ 更值得关注的是，根据我们的流程，1BW与自身训练集的重叠率高达近13.2%。

在Winograd模式挑战中，我们仅发现10个模式与WebText训练集存在8元词重叠。其中2个为虚假匹配。剩余8个模式中，仅有1个出现在任何上下文中。

⁴额外重叠内容的显著部分源于编辑者在多个主题相近的文章（如朝鲜战争中不同战役的描述）中重复使用某些段落。

泄露了答案。

在CoQA数据集上，新闻领域约15%的文档已存在于WebText中，模型在这些文档上F1值提升约3分。CoQA开发集指标基于5个不同领域的平均表现，我们测得跨领域重叠效应带来约0.5-1.0的F1值提升。但需注意：由于CoQA发布时间晚于WebText链接截止日期，WebText中不包含任何实际训练问答数据。

在LAMBADA测试中，平均重叠率为1.2%。当重叠率超过15%时，GPT-2的困惑度表现提升约2个单位。当排除所有存在重叠的样本后，重新计算指标导致困惑度从8.6升至8.7，准确率从63.2%降至62.9%。整体结果的微小变化源于仅有1/200的样本存在显著重叠。

总体而言，我们的分析表明WebText训练数据与特定评估数据集之间的数据重叠，为报告结果带来微小但稳定的提升。然而，如表6所示，对于多数数据集，我们并未观察到显著大于标准训练集与测试集之间重叠程度的现象。

理解并量化高度相似文本对性能的影响是重要研究课题。采用可扩展模糊匹配等更优去重技术亦有助于解答这些问题。目前我们建议在新建NLP数据集时，将基于n-gram重叠度的去重技术作为训练集与测试集划分过程中的关键验证步骤和合理性检查。

判断WebText语言模型性能是否源于记忆的另一种方法，是考察其在保留测试集上的表现。如图4所示，WebText在训练集和测试集上的表现趋同，且随模型规模增大同步提升。这表明即使GPT-2在WebText数据集上仍存在多方面的欠拟合现象。

GPT-2还能撰写关于会说话独角兽发现的新闻报道。具体示例见表13。

5. 相关研究

本研究的重要部分评估了基于大规模数据集训练的大型语言模型的性能。

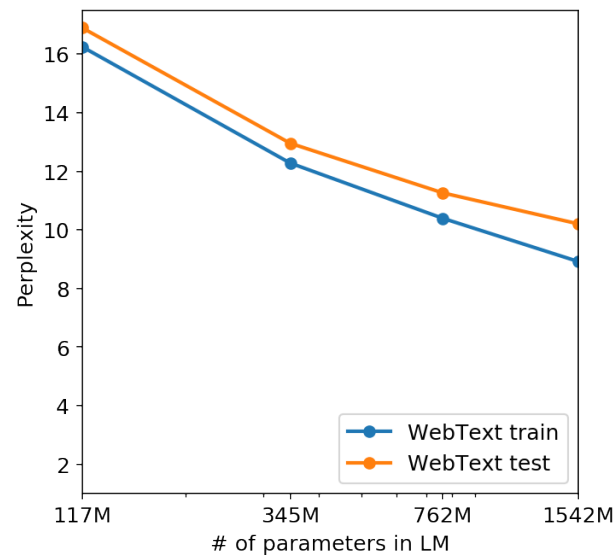


Figure 4. The performance of LMs trained on WebText as a function of model size.

is similar to the work of Jozefowicz et al. (2016) which scaled RNN based language models on the 1 Billion Word Benchmark. Bajjar 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

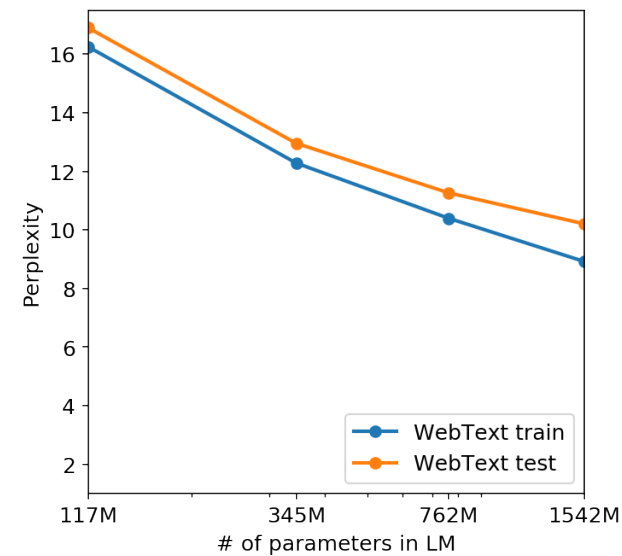


图4. 基于WebText训练的语言模型性能随模型规模的变化曲线。

这与Jozefowicz等人（2016）的工作相似，该研究在10亿词基准上扩展了基于RNN的语言模型。Bajjar等人（2016）此前也通过从古腾堡计划创建更庞大的训练数据集来补充标准训练数据集，从而在儿童图书测试中提升了结果。Hestness等人（2017）深入分析了深度学习模型性能随模型容量与数据集规模变化的函数关系。尽管我们的实验在不同任务间存在较大噪声，但结果表明：目标任务的子任务中同样存在类似趋势，且该趋势延续至10亿参数（1B+）规模范畴。

生成模型中已记录过有趣的学习功能，例如 Karpathy等人（2015）提出的RNN语言模型单元可执行行宽追踪与引号/评论检测。对本研究更具启发性的是Liu等人（2018）的发现：训练生成维基百科文章的模型同时习得了跨语言名称转换能力。

先前研究已探索过滤与构建大型网页语料库的替代方案，例如iWeb语料库（Davies, 2018）。

针对语言任务的预训练方法已有大量研究。除引言中提及的方法外，GloVe (Pennington等人，2014) 将词向量表示学习扩展至整个Common Crawl数据集。文本深度表征学习的早期代表性工作是跳跃式思考向量（Skip-thought Vectors, Kiros等人，2015）。McCann等人（2017）探索了机器翻译模型衍生表征的应用，而Howard与Ruder（2018）则

改进(Dai & Le, 2015)提出的基于循环神经网络的微调方法。(Conneau et al., 2017a)研究了自然语言推理模型学习到的表示迁移性能，(Subramanian et al., 2018)则探索了大规模多任务训练。

(Ramachandran et al., 2016) 证明序列到序列模型通过预训练语言模型作为编码器和解码器进行初始化能获得显著提升。最新研究表明，语言模型预训练同样有助于优化复杂生成任务（如闲聊对话和基于对话的问答系统）的微调效果（Wolf et al., 2019）（Dinan et al., 2018）。

6. 讨论

大量研究致力于学习（Hill et al., 2016）、理解（Levy & Goldberg, 2014）和批判性评估（Wieting & Kiela, 2019）监督式与非监督式预训练方法的表示形式。我们的研究结果表明，无监督任务学习是值得探索的全新研究方向。这些发现可能有助于解释预训练技术在下游NLP任务中取得广泛成功的原因——我们证明在极限情况下，其中一种预训练技术开始直接学习执行任务，无需监督式调整或修改。

在阅读理解任务中，GPT-2在零样本环境下的表现可与监督学习基线媲美。但在摘要生成等其他任务上，尽管其能完成任务，但定量指标显示其表现仍处于初级阶段。虽然该研究结果具有启示意义，但就实际应用而言，GPT-2的零样本性能仍远未达到可用水平。

我们已研究WebText语言模型在众多经典NLP任务上的零样本表现，但仍有大量可评估的附加任务。在许多实际场景中，GPT-2的表现仍不优于随机猜测。即便在我们评估的常见任务（如问答和翻译）中，语言模型也需具备足够容量才能开始超越基础基线。

虽然零样本性能为GPT-2在众多任务中的潜在表现建立了基准，但其通过微调能达到的上限尚不明确。在某些任务中，GPT-2完全抽象化的输出与基于提取式指针网络（Vinyals等人，2015）的输出存在显著差异——后者目前仍是许多问答和阅读理解数据集上的顶尖方案。鉴于GPT模型在微调方面的既往成功，我们计划在deca NLP和GLUE等基准测试中探索微调方案，尤其需要验证额外训练是否能提升模型在这些任务中的表现。

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.

⁵Preliminary 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.

GPT-2的训练数据量与模型容量足以克服BERT（Devlin et al., 2018）所展现的单向表示效率不足问题。

7. 结论

当大型语言模型在足够庞大且多样化的数据集上训练后，便能横跨多个领域和数据集展现优异性能。GPT-2在8个测试语言建模数据集中，有7个实现了零样本训练下的最先进性能。该模型在零样本环境下展现的任务多样性表明：经过高容量训练以最大化多样化文本语料概率的模型，开始能够在无需显式监督的情况下执行大量任务。⁵

鸣谢

感谢所有撰写文本、分享链接并在WebText平台为内容点赞的用户。数百万参与者共同创建了GPT-2的训练数据集。同时感谢所有协助我们构建训练基础设施的谷歌员工，包括Zak Stone、JS Riehl、Jonathan Hseu、Russell Power、Youlong Cheng、Noam Shazeer、Solomon Boulos、Michael Banfield、Aman Gupta、Daniel Sohn等众多同仁。最后感谢为论文初稿提供反馈的各位：Jacob Steinhardt、Sam Bowman、Geoffrey Irving和Madison May。

参考文献

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年。

Alcorn, M., Nguyen, A., Gong, Z., Wang, C., Mai, L., Ku, W.-S., and Nguyen, A. 摆出姿势王,神经网络易受熟悉物体怪异姿势的欺骗. *arXiv预印本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: 英语与汉语端到端语音识别. 发表于国际机器学习会议, 第173–182页, 2016.

Artetxe, M., Labaka, G., and Agirre, E. Unsupervised neural machine translation. *arXiv预印本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小型模型下载及使用初始代码可访问：<https://github.com/openai/gpt-2>

Ba, J. L., Kiros, J. R., and Hinton, G. E. 层归一化. *arXiv预印本 arXiv:1607.06450*, 2016. Bajgar, O., Kadlec, R., and Kleindienst, J. 拥抱数据丰裕：阅读理解的Booktest数据集. *arXiv预印本arXiv:1610.00956*, 2016. Barz, B. and Denzler, J. 我们在测试数据上训练吗？清除CIFAR中的近似重复数据. *arXiv预印本arXiv:1902.00423*, 2019. Bengio, Y., Ducharme, R., Vincent, P., 和 Jauvin, C. 神经概率语言模型. *机器学习研究期刊*, 3(2月):1137–1155, 2003. Bowman, S. R., Pavlick, E., Grave, E., Van Durme, B., 王, A., 胡拉, J., 夏, P., 帕帕加里, R., 麦考伊, R. T., 帕特尔, R., 等. 寻找艾尔莫的朋友：超越语言建模的句子级预训练. *arXiv预印本arXiv:1812.10860*, 2018.

Caruana, R. 多任务学习. *机器学习*, 28 75, 1997 1 41(): – .

切尔巴, C., Koehn, P., and Robinson, T. 统计语言建模进展评估的十亿词基准测试. *arXiv预印本 arXiv:1312.3005*, 2013.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. 自然语言处理（几乎）从零开始. *机器学习研究期刊*, 12(8月):2493–2537, 2011.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A. 《基于自然语言推理数据的通用句子表示监督学习》. *arXiv预印本 arXiv:1705.02364*, 2017a.

Conneau, A., and Jégou, H. 无平行数据的单词翻译. *arXiv预印本 arXiv:1710.04087*, 2017b。

Dai, A. M. 以及Le, Q. V. 半监督序列学习. 载于《神经信息处理系统进展》, 第3079–3087页, 2015年。

Dai, Z. , Yang, Z. , and Salakhutdinov, R. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

戴维斯, M. 140亿词规模的iweb语料库. *https://corpus.byu.edu/iWeb/*, 2018.

Dehghani, M., and Kaiser, Ł. Universal transformers. *arXiv preprint arXiv:1807.03819*, 2018. .

Devlin, J. Bert: 深度双向变换器的预训练及其在语言理解中的应用. *arXiv预印本 arXiv:1810.04805*, 2018.

Dinan, E., 以及Weston, J. 《维基百科的魔法师：知识驱动的对话代理》. *arXiv预印本arXiv:1811.01241*, 2018.

Fan, A., Lewis, M., and Dauphin, Y. 层级神经故事生成模型. *arXiv预印本arXiv:1805.04833*,2018.

Language Models are Unsupervised Multitask Learners	
Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. <i>arXiv preprint arXiv:1703.03400</i> , 2017.	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. <i>Proceedings of the national academy of sciences</i> , pp. 201611835, 2017.
Gehrmann, S., Deng, Y., and Rush, A. M. Bottom-up abstractive summarization. <i>arXiv preprint arXiv:1808.10792</i> , 2018.	Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., and Fidler, S. Skip-thought vectors. In <i>Advances in neural information processing systems</i> , pp. 3294–3302, 2015.
Gillick, D., Brunk, C., Vinyals, O., and Subramanya, A. Multilingual language processing from bytes. <i>arXiv preprint arXiv:1512.00103</i> , 2015.	Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In <i>Advances in neural information processing systems</i> , pp. 1097–1105, 2012.
Gong, C., He, D., Tan, X., Qin, T., Wang, L., and Liu, T.-Y. Frage: frequency-agnostic word representation. In <i>Advances in Neural Information Processing Systems</i> , pp. 1341–1352, 2018.	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.
Grave, E., Joulin, A., and Usunier, N. Improving neural language models with a continuous cache. <i>arXiv preprint arXiv:1612.04426</i> , 2016.	Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. Building machines that learn and think like people. <i>Behavioral and Brain Sciences</i> , 40, 2017.
He, K., Zhang, X., Ren, S., and Sun, J. Identity mappings in deep residual networks. In <i>European conference on computer vision</i> , pp. 630–645. Springer, 2016.	Lample, G., Conneau, A., Denoyer, L., and Ranzato, M. Unsupervised machine translation using monolingual corpora only. <i>arXiv preprint arXiv:1711.00043</i> , 2017.
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. <i>arXiv preprint arXiv:1712.00409</i> , 2017.	Levesque, H., Davis, E., and Morgenstern, L. The winograd schema challenge. In <i>Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning</i> , 2012.
Hill, F., Bordes, A., Chopra, S., and Weston, J. The goldilocks principle: Reading children’s books with explicit memory representations. <i>arXiv preprint arXiv:1511.02301</i> , 2015.	Levy, O. and Goldberg, Y. Neural word embedding as implicit matrix factorization. In <i>Advances in neural information processing systems</i> , pp. 2177–2185, 2014.
Hill, F., Cho, K., and Korhonen, A. Learning distributed representations of sentences from unlabelled data. <i>arXiv preprint arXiv:1602.03483</i> , 2016.	Liu, P. J., Saleh, M., Pot, E., Goodrich, B., Sepassi, R., Kaiser, L., and Shazeer, N. Generating wikipedia by summarizing long sequences. <i>arXiv preprint arXiv:1801.10198</i> , 2018.
Hoang, L., Wiseman, S., and Rush, A. M. Entity tracking improves cloze-style reading comprehension. <i>arXiv preprint arXiv:1810.02891</i> , 2018.	McCann, B., Bradbury, J., Xiong, C., and Socher, R. Learned in translation: Contextualized word vectors. In <i>Advances in Neural Information Processing Systems</i> , pp. 6294–6305, 2017.
Howard, J. and Ruder, S. Universal language model fine-tuning for text classification. In <i>Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , volume 1, pp. 328–339, 2018.	McCann, B., Keskar, N. S., Xiong, C., and Socher, R. The natural language decathlon: Multitask learning as question answering. <i>arXiv preprint arXiv:1806.08730</i> , 2018.
Jelinek, F. and Mercer, R. L. Interpolated estimation of markov source parameters from sparse data. In <i>Proceedings of the Workshop on Pattern Recognition in Practice, Amsterdam, The Netherlands: North-Holland, May.</i> , 1980.	Merity, S., Xiong, C., Bradbury, J., and Socher, R. Pointer sentinel mixture models. <i>arXiv preprint arXiv:1609.07843</i> , 2016.
Jia, R. and Liang, P. Adversarial examples for evaluating reading comprehension systems. <i>arXiv preprint arXiv:1707.07328</i> , 2017.	Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. In <i>Advances in neural information processing systems</i> , pp. 3111–3119, 2013.
Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., and Wu, Y. Exploring the limits of language modeling. <i>arXiv preprint arXiv:1602.02410</i> , 2016.	Nallapati, R., Zhou, B., Gulcehre, C., Xiang, B., et al. Abstractive text summarization using sequence-to-sequence rnns and beyond. <i>arXiv preprint arXiv:1602.06023</i> , 2016.
Kaiser, L., Gomez, A. N., Shazeer, N., Vaswani, A., Parmar, N., Jones, L., and Uszkoreit, J. One model to learn them all. <i>arXiv preprint arXiv:1706.05137</i> , 2017.	Paperno, D., Kruszewski, G., Lazaridou, A., Pham, Q. N., Bernardi, R., Pezzelle, S., Baroni, M., Boleda, G., and Fernández, R. The lambda dataset: Word prediction requiring a broad discourse context. <i>arXiv preprint arXiv:1606.06031</i> , 2016.
Karpathy, A., Johnson, J., and Fei-Fei, L. Visualizing and understanding recurrent networks. <i>arXiv preprint arXiv:1506.02078</i> , 2015.	Pennington, J., Socher, R., and Manning, C. Glove: Global vectors for word representation. In <i>Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)</i> , pp. 1532–1543, 2014.

语言模型是无监督的多任务学习者		
Finn, C., Abbeel, P., and Levine, S. 《用于深度网络快速适应的模型无关元学习》. arXiv预印本 arXiv:1703.03400, 2017.	Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. 突破神经网络中的灾难性遗忘问题。《美国国家科学院院刊》，第201611835页，2017年。	
Gehrmann, S., Deng, Y., and Rush, A. M. 自下而上抽象化摘要生成. arXiv预印本 arXiv:1808.10792, 2018年。	Kiros, R., Torralba, A., and Fidler, S. 跳跃思考向量. 载于《神经信息处理系统进展》，第3294–3302页，2015年。	
Gillick, D., Brunk, C., Vinyals, O., 以及Subramanya, A. 多语言处理技术研究（by bytes）。arXiv预印本 arXiv:1512.00103, 2015。	Krizhevsky, A. 基于深度卷积神经网络的ImageNet分类。载于《神经信息处理系统进展》，第1097–1105页，2012年。	
Gong, C 问题：频率无关词表征。载于《神经信息处理系统进展》，第1341–1352页，2018年。	Kwiatkowski, T., Palomaki, J., Rhinehart, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Kelcey, M., Devlin, J., et al. 自然问题：问答研究基准测试. 2019.	
Grave, E 《基于连续缓存改进神经网络语言模型》。arXiv预印本 arXiv:1612.04426, 2016年。	Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. 《构建具备人类般学习与思考能力的机器》。《行为与脑科学》，40卷，2017年。Lample, G., Conneau, A., Denoyer, L., and Ranzato, M. 仅使用单语语料库的无监督机器翻译. arXiv预印本 arXiv:1711.00043, 2017. Levesque, H., Davis, E., and Morgenstern, L. 温诺格拉德模式挑战. 载于第十三届知识表示与推理原理国际会议论文集，2012年。Levy, O. 与 Goldberg, Y. 神经词嵌入作为隐式矩阵分解。载于神经信息处理系统进展论文集，第2177–2185页，2014年。Liu, P. J., Saleh, M., Pot, E., Goodrich, B., Sepassi, R., Kaiser, L., and Shazeer, N. 通过长序列摘要生成维基百科. arXiv预印本 arXiv:1801.10198, 2018. McCann, B., Bradbury, J., Xiong, C., and Socher, R. 翻译中的学习：语境化词向量。收录于《神经信息处理系统进展》，第6294–6305页，2017年。McCann, B., Keskar, N. S., Xiong, C., and Socher, R. 自然语言十项全能：作为问答的多任务学习。arXiv预印本arXiv:1806.08730, 2018年。Merity, S., Xiong, C., Bradbury, J., and Socher, R. 指针哨兵混合模型。arXiv预印本arXiv:1609.07843, 2016. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. 词与短语的分布式表示及其组合性. 收录于《神经信息处理系统进展》，第3111–3119页，2013. Nallapati, R., Zhou, B., Gulcehre, C., Xiang, B., 等. 基于序列到序列RNN的抽象文本摘要及其拓展. arXiv预印本arXiv:1602.06023, 2016. Paperno, D., Kruszewski, G., Lazaridou, A., Pham, Q. N., Bernardi, R., Pezzelle, S., Baroni, M., Boleda, G., and Fernández, R. 朗巴达数据集：需要广泛语篇背景的词预测. arXiv预印本arXiv:1606.06031, 2016. Pennington, J., Socher, R., and Manning, C. GLOVE：词表示的全球向量。收录于《2014年自然语言处理经验方法会议论文集》（EMNLP），第1532–1543页，2014年。	He, K., Zhang, X., Ren, S., and Sun, J. 深度残差网络中的恒等映射. 载于欧洲计算机视觉会议论文集, 第630–645页。Springer出版社, 2016. Hestness, J., Narang, S., Ardalani, N., 迪亚莫斯, G., 俊, H., 基安-伊内贾德, H., 帕特瓦里, M., 阿里, M., 杨, Y., 周, Y. 深度学习的可扩展性具有可预测性（实证研究）。arXiv预印本 arXiv:1712.00409, 2017年。希尔, F., 博尔德斯, A., 乔普拉, S., 和韦斯顿, J. 金发姑娘原则：通过显式记忆表征阅读儿童读物. arXiv预印本arXiv:1511.02301, 2015年。希尔, F., 赵, K., 和科霍宁, A. 从无标签数据中学习句子的分布式表征。arXiv预印本arXiv:1602.03483, 2016年。Hoang, L., Wiseman, S., and Rush, A. M. 实体追踪提升填空式阅读理解. arXiv预印本arXiv:1810.02891, 2018年。Howard, J. 和 Ruder, S. 文本分类的通用语言模型微调。载于第 56 届计算语言学协会年会论文集（第 1 卷：长论文），第 1 卷，第 328-339 页，2018 年。Jelinek, F. 和 Mercer, R. L. 基于稀疏数据的马尔可夫源参数插值估计。载于实践模式识别研讨会论文集，荷兰阿姆斯特丹：北荷兰出版社，1980年5月。贾瑞、梁鹏. 用于评估阅读理解系统的对抗样本. arXiv 预印本 arXiv:170707328, 2017. Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., and Wu, Y. 语言建模极限探索. arXiv预印本arXiv:1602.02410, 2016. Kaiser, L., Gomez, A. N., Shazeer, N., Vaswani, A., Parmar, N., Jones, L., 和乌兹科雷特, J. 学习所有模型的单一模型。arXiv预印本arXiv:1706.05137, 2017年。卡帕西, A., 约翰逊, J., 和李飞飞. 可视化与理解循环网络. arXiv预印本 arXiv:150602078, 2015年。

Language Models are Unsupervised Multitask Learners	
Peters, M. E. and Lecocq, D. Content extraction using diverse feature sets. In <i>Proceedings of the 22nd International Conference on World Wide Web</i> , pp. 89–90. ACM, 2013.	Vinyals, O., Fortunato, M., and Jaitly, N. Pointer networks. In <i>Advances in Neural Information Processing Systems</i> , pp. 2692–2700, 2015.
Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. Deep contextualized word representations. <i>arXiv preprint arXiv:1802.05365</i> , 2018.	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. <i>arXiv preprint arXiv:1804.07461</i> , 2018.
Radford, A., Jozefowicz, R., and Sutskever, I. Learning to generate reviews and discovering sentiment. <i>arXiv preprint arXiv:1704.01444</i> , 2017.	Weston, J. E. Dialog-based language learning. In <i>Advances in Neural Information Processing Systems</i> , pp. 829–837, 2016.
Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. Improving language understanding by generative pre-training. 2018.	Wieting, J. and Kiela, D. No training required: Exploring random encoders for sentence classification. <i>arXiv preprint arXiv:1901.10444</i> , 2019.
Ramachandran, P., Liu, P. J., and Le, Q. V. Unsupervised pre-training for sequence to sequence learning. <i>arXiv preprint arXiv:1611.02683</i> , 2016.	Wolf, T., Sanh, V., Chaumond, J., and Delangue, C. Transfer-transfo: A transfer learning approach for neural network based conversational agents. <i>arXiv preprint arXiv:1901.08149</i> , 2019.
Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do cifar-10 classifiers generalize to cifar-10? <i>arXiv preprint arXiv:1806.00451</i> , 2018.	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. <i>arXiv preprint arXiv:1901.11373</i> , 2019.
Reddy, S., Chen, D., and Manning, C. D. Coqa: A conversational question answering challenge. <i>arXiv preprint arXiv:1808.07042</i> , 2018.	
Schwartz, R., Sap, M., Konstas, I., Zilles, L., Choi, Y., and Smith, N. A. Story cloze task: Uw nlp system. In <i>Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics</i> , pp. 52–55, 2017.	
See, A., Liu, P. J., and Manning, C. D. Get to the point: Summarization with pointer-generator networks. <i>arXiv preprint arXiv:1704.04368</i> , 2017.	
Sennrich, R., Haddow, B., and Birch, A. Neural machine translation of rare words with subword units. <i>arXiv preprint arXiv:1508.07909</i> , 2015.	
Subramanian, S., Trischler, A., Bengio, Y., and Pal, C. J. Learning general purpose distributed sentence representations via large scale multi-task learning. <i>arXiv preprint arXiv:1804.00079</i> , 2018.	
Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks. In <i>Advances in neural information processing systems</i> , pp. 3104–3112, 2014.	
Sutskever, I., Jozefowicz, R., Gregor, K., Rezende, D., Lillicrap, T., and Vinyals, O. Towards principled unsupervised learning. <i>arXiv preprint arXiv:1511.06440</i> , 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. <i>arXiv preprint arXiv:1811.01778</i> , 2018.	
Trinh, T. H. and Le, Q. V. A simple method for commonsense reasoning. <i>arXiv preprint arXiv:1806.02847</i> , 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 <i>Advances in Neural Information Processing Systems</i> , pp. 5998–6008, 2017.	
Vinyals, O. and Le, Q. A neural conversational model. <i>arXiv preprint arXiv:1506.05869</i> , 2015.	

语言模型是无监督的多任务学习者	
Peters, M. E. 与 Lecocq, D. 《基于多样化特征集的内容提取》。载于《第22届国际万维网会议论文集》，第89–90页。ACM, 2013年。Peters, M. E.、Neumann, M.、Iyyer, M.、Gardner, M.、Clark, C.、Lee, K. 及 Zettlemoyer, L. 深度上下文词表示. arXiv预印本 arXiv:1802.05365, 2018. Radford, A., Jozefowicz, R., 及 Sutskever, I. 学习生成评论与情感发现. arXiv预印本 arXiv:1704.01444, 2017. 拉德福德, A., 纳拉辛汉, K., 萨利曼斯, T., 及苏茨克弗, I. 通过生成式预训练提升语言理解能力。2018年。拉马钱德兰, P., 刘, P. J., 及黎, Q. V. 序列到序列学习的无监督预训练. arXiv 预印本 arXiv:1611.02683, 2016 年。雷希特, B., 罗洛夫斯, R., 施密特, L., 尚卡尔, V. CIFAR-10分类器能否泛化到CIFAR-10? arXiv预印本arXiv:1806.00451, 2018年。Reddy, S., Chen, D., and Manning, C. D. Coqa: 对话式问答挑战赛. arXiv预印本 arXiv:1808 07042, 2018. Schwartz, R., Sap, M., 康斯塔斯, I., 齐尔斯, L., 崔, Y., 及史密斯, N. A. 故事填空任务: UW NLP系统. 载于《第二届词汇、句子及话语层级语义关联模型研讨会论文集》, 第52–55页, 2017. See, A., Liu, P. J., and Manning, C. D. 直击要点: 基于指针生成网络的摘要生成. arXiv预印本 arXiv:1704.04368, 2017. Sennrich, R., 哈多, B., 伯奇, A. 基于亚词单元的罕见词神经机器翻译. arXiv预印本 arXiv:1508.07909, 2015年。苏布拉马尼安, S., 特里施勒, A., 本吉奥, Y., 帕尔, C. J. 通过大规模多任务学习获取通用分布式句子表征。 arXiv预印本arXiv:180400079, 2018. Sutskever, I., Vinyals, O., and Le, Q. V. 基于神经网络的序列到序列学习. 载于《神经信息处理系统进展》, 第3104–3112页, 2014. Sutskever, I., Jozefowicz, R., Gregor, K., Rezende, D., Lillicrap, T., and Vinyals, O. 迈向原理性无监督学习. arXiv预印本 arXiv:1511.06440, 2015. Trichelair, P., Emami, A., Cheung, J. C. K., Trischler, A., Suleman, K., and Diaz, F. 论自然语言理解中常识推理的评估. arXiv预印本 arXiv:1811.01778, 2018. Trinh, T. H. 和 Le, Q. V. 常识推理的简易方法. arXiv预印本 arXiv:1806.02847, 2018. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., 戈麦斯, A. N., 凯泽, Ł., 波洛苏金, I. 注意力机制即解决方案。载于《神经信息处理系统进展》, 第5998–6008页, 2017年。维尼尔斯, O. 与黎, Q. 神经对话模型. arXiv预印本arXiv:1506.05869, 2015年。	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: 自然语言理解的多任务基准与分析平台。arXiv预印本arXiv:1804.07461, 2018. Weston, J. E. 对话式语言学习。载于《神经信息处理系统进展》, 第 829–837 页, 2016 年。维廷, J.与基拉, D. 无需训练: 探索随机编码器在句子分类中的应用。arXiv预印本arXiv:1901.10444, 2019年。Wolf, T., Sanh, V., Chaumond, J., and Delangue, C. 迁移转换器: 基于神经网络的对话代理迁移学习方法. arXiv预印本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., 等. 学习与评估通用语言智能. arXiv预印本 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.

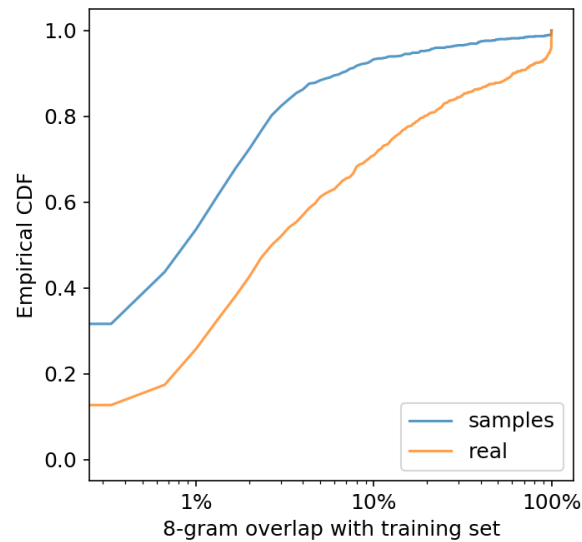


Figure 5. CDF of percentage 8-gram overlap with WebText training set, for both WebText test set and samples (conditioned on WebText test set, with top- k truncated random sampling with $k = 40$). Most samples have less than 1% overlap, including over 30% of samples with no overlap, whereas the median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

8. 附录A：样例

8.1. 模型容量

为补充图4所示大型语言模型在WebText数据集上实现的困惑度提升，表7至表11展示了最小规模WebText语言模型与GPT-2在随机未见WebText测试集文章上的并列补全结果。

8.2. 文本记忆

我们观察到GPT-2在处理数据集中多次重复的长字符串（如名言或演讲）时存在记忆行为。例如，当以《葛底斯堡演说》开篇一句半（在WebText中约出现40次）作为条件时，GPT-2通过argmax解码可完整复现该演说。即使在无截断采样情况下，我们发现模型在漂移前仍会持续复制演说内容，尽管风格趋同。通常漂移过程在100-200个词元内完成，漂移后则呈现日益扩大的多样性。

为量化样本中精确记忆的出现频率，我们基于WebText测试集文章对GPT-2进行条件训练生成样本，并比较GPT-2生成文本与真实答案的重合率。分析结果如下所示，表明GPT-2重复训练集文本的频率低于预留文章的基准率。

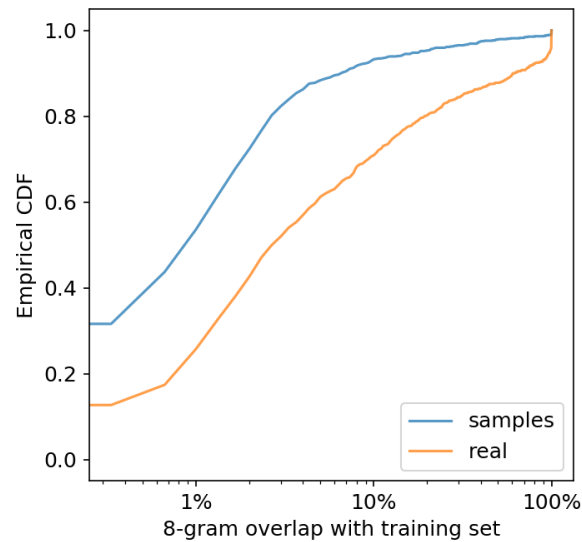


图5. WebText测试集与样本数据的8-gram重叠率累积分布曲线（基于WebText测试集条件，采用 k 截断随机采样与 $k = 40$ ）。多数样本重叠率低于1%（其中30%以上样本无重叠），而测试集中位数重叠率为2.6%。

8.3. 多样性

表12展示了同一随机WebText测试集上下文的多重完成结果，呈现了标准采样设置下完成结果的多样性。

8.4. 鲁棒性

表13展示了前文提及的会说话独角兽新闻报道。我们发现该模型能够处理分布外语境，但此类样本的质量普遍较低。

Context (WebText test)	
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	
Small model completion	GPT-2 completion
.	for me!
Outfit 2: High End	Outfit 2: Casual Jacket
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!	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: High Light	Outfit 3: Dress
This was a ”high light” look while I had a chance to get some color. I think the color is based on	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

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.

上下文（WebText测试）	
配置1：典型	
这双鞋买回来时，我首先想到的就是搭配这套衣服。这就像是它的夏季版本杰克·格兰瑟姆的着装风格；事实上，我的鞋子颜色与他的耐克竞速鞋极为接近！与厚重的哈里斯粗花呢相比夹克与牛仔衬衫，我正穿着棉质双排扣夹克和亚麻衬衫。这两种面料（在此配色下）都是夏季必备单品，因其既可搭配深色长裤，亦能与浅色长裤相得益彰！如您所见，它们堪称绝配。搭配深色牛仔裤和鞋子。这是种伪男装/学院风的穿搭。	
整体而言这是非常休闲的装扮，因此我搭配了运动鞋。我不赞成像《GQ》杂志常展示的那样，穿着全羊毛西装搭配运动鞋，但我绝对支持保持休闲风格，比如这件棉质双排扣外套。休闲面料是将运动鞋融入休闲男装的关键。我甚至建议搭配卡其色斜纹棉布西装或白色亚麻西装穿着运动鞋。只需确保不系领带，或搭配T恤或Polo衫；系领带配运动鞋会显得过于正式。	
小模型补全	GPT-2 生成
.	为我！
套装2：高端款	套装2：休闲夹克
这套服装恰好处于风暴的中心。融合ASICS专业运动装备与街头服饰（如J. Crew）的混合风格以及J. Crew如何处理那些略带休闲风格的单品更传统。重申之前提到的，这是低预算项目可能极其昂贵。这些将若试图以兼职身份工作，这绝对是个问题sneakermaker，这没问题，但若您想尝试他们服装的低价版本！	我明白有些人可能反对佩戴穿着高领衫配运动鞋，但我想要忠于自我，因此我选择了修身高领衫，颜色更偏向...类似于我今日最爱的颜色。经典高领衫，款式不过分张扬。我确实换了件为这些亚瑟士Tiger Killshots运动鞋配鞋，真希望当初没这么做已完成。这双鞋本身非常舒适，然而，我发现穿着这些运动鞋时，我的脚部看起来略大。我会搭配牛仔裤、斜纹棉布裤穿着。随你便，但绝不建议穿它们仅凭此无法确定。涉及的变量实在太多。
配置3：高亮模式	套装3：连衣裙
这是我偶然获得的"亮点"一瞥添加一些色彩。我认为色彩基于	这是我最后穿的这套衣服致命一击。此版本的ASICS在性能上有了显著提升。现代精简版。这双鞋拥有最这是我能想到最现代简约的设计！包含大量黑色款运动鞋造型酷炫，绝对能非常适合夏季穿着。
	对我而言，搭配西装穿着是最简便的方式拥有一双优质运动鞋。这基本就是它！

表7. 随机未见上下文（顶部），以及最小（左）与最大（右）模型生成的非精选补全结果。上下文为768个令牌，展示约256令牌的段落内容。补全结果为256个令牌并完整呈现。生成采用顶部 k 随机采样结合 $k = 40$ 模型。

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.	<p>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.</p> <p>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.</p> <p>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.</p> <p>The last Muslim state in Spain, Granada, was conquered by the</p>

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.

上下文（WebText测试）	
西班牙部分最辉煌的历史景点可追溯至穆斯林统治时期， 包括..... 梅斯基塔大教堂最初作为科尔多瓦大清真寺建造， 与同在科尔多瓦的麦地那·阿扎哈拉相邻， 现已成废墟但至今仍可参观的麦地那·萨赫拉（Madinat al-Zahra）——安达卢西亚宫殿； 以及格拉纳达的阿罕布拉宫（Alhambra）， 一座宏伟完好的宫殿。 另有两座犹太会堂仍屹立不倒， 它们建于穆斯林统治时期的西班牙：托莱多的圣玛丽亚·拉布兰卡教堂与科尔多瓦老城区的犹太教堂。	
收复失地运动与帝国时代 [编辑]	
小模型补全	GPT-2 生成
穆斯林统治时期的西班牙与 世界其他地区，拥有相当庞大的中世纪历史。 尽管西班牙帝国存在于公元前9世纪后半叶， 其发展相当缓慢。到公元1150年， 该地区已转变为一个小型世俗化帝国。处于更为繁荣多元的状态。公元1050年那不勒斯皇帝， 至少部分特征体现在一系列战争与入侵始于西北沿海地区至该地区西海岸（例如多瑙河沿岸）河流自直布罗陀延伸至西班牙南部， 但很快就在同一天被摧毁了罗马军队占领日）， 自称皇帝那不勒斯。他在过去几年中做出了重大贡献。耗费数十年重建并维系帝国， 首个其继任者是费迪南德， 他1148年从教皇宝座夺取权力， 而首位帝国继任者是瑞典国王阿尔弗雷德二世。帝国再次成为少数几个独立国家之一在欧洲大陆， 尽管人数不超过200百万人口在此期间居住于其领土之上人们向皇帝陈述了许多事情。从1720年起穆斯林曾占西班牙人口的一半。	"收复失地运动"（字面意为"重返故土"） 1092年至1095年间，伊比利亚半岛建立了卡斯蒂利亚与阿拉贡联合王国的半岛穆斯林王朝将存在。这是最后一个欧洲时期穆斯林统治将被视为"native"以及最后一次由Christian引发的事件有组织的冲突将显现于欧洲。 穆斯林统治时期 西班牙的摩尔人常被重新——被提及的，尤其在非穆斯林群体中， 作为"格拉纳达征服"与西班牙有时被称为将"Granada"译为西班牙的"格拉纳达"。穆斯林统治的最后岁月随后是西班牙的收复失地运动， 其中穆斯林被驱逐或被迫离开伊比利亚半岛半岛战役中费迪南德九世取得胜利1189年3月普瓦捷战役及耶路撒冷攻陷事件公元1187年由萨拉丁攻陷。 西班牙的收复失地运动时期宣告终结当摩尔人撤退至北方时。然而摩尔人的基督教规则将取代该规则直至当收复失地运动持续期间，收复失地运动时代得以延续从1229年至1492年。 西班牙最后的穆斯林国家， 格拉纳达，曾是——由

表8. 随机未见上下文（顶部）， 以及最小（左）与最大（右）模型生成的非筛选补全结果。上下文为768个令牌， 展示段落约256个令牌。补全结果为256个令牌并完整呈现。生成采用顶部*k* 随机采样结合 *k* = 40 模型。

Language Models are Unsupervised Multitask Learners		
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 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	GPT-2 completion 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.

语言模型是无监督的多任务学习者		
上下文（WebText测试）		
Pinky		
粉色幽灵的人工智能设计旨在"感知"与红色幽灵行为相反的模式。Pinky实际上试图在吃豆人前方布设目标。具体实现方式是将目标设置在吃豆人当前位置前方4格处。吃豆人移动的方向。唯一例外是当吃豆人向上移动时。由于溢出代码存在错误，计算中包含的左偏移量等于预期上偏移量。		
利用吃豆人（Pac-Man）的预期位置，玩家可利用粉豆（Pinky）的导航代码。这在常见于吃豆人与粉豆人彼此邻近的场景。例如当它们当双方距离小于4格时，目标格现在位于Pinky身后。若他遭遇在此配置下，当行进至交叉路口时，他将尝试通过转向折返以抵达目标方块。		
玩家常通过快速变向实施所谓的"假动作"——当Pinky接近十字路口时。通过在Pinky移动目标方块的瞬间		
小模型补全 若从左侧接近，他们将获得"正确"的方向。 若实际位置为远离what Pac-当玩家处于该状态时，角色将不会对幽灵的指令作出响应。移动，幽灵将转而返回玩家的当前位置。这是因为吃豆人不会移动通过转译。 幽灵将不得不前往另一个地点某种方式使他得以回归起点，而这正是实现许多方面。它通常会直接保留位置，例如例如如果吃豆人真的从悬崖跳下，他会返回他消失前所在的位置空格。这些幽灵存在相当多的可能性。路径或许存在，但最显而易见的当属玩家希望避免陷入游戏之中，目标简单如让吃豆人留在游戏中1/3 那么长。 追逐 以吃豆人为例，这类似于吃豆人能在战斗中"迎头赶上"。 幽灵能穿透敌人追击。此为使用——出于诸多原因，无论是对吃豆人还是其人工智能而言，这都至关重要。与多数幽灵游戏不同，这个幽灵仅在方向来自	GPT-2 生成 正要穿越十字路口时，吃豆人就会能够躲避幽灵的弹幕并返回安全区域安全屋的位置。 Buster 黄色ghost的人工智能系统专门旨在尽量避免吃豆人离开安全屋。玩家必须按空格键才能显示安全区域房屋的门，然后走进门内寻找黄色幽灵藏在隔壁房间。 玩家还需注意幽灵不会若发现吃豆人已出现，则尝试逃离安全屋事实上，幽灵在最初阶段根本不会尝试逃离。除非他们已身处安全屋。 这两种行为基于以下事实：玩家并非总是试图抵达安全屋在吃豆人（或闪光点）之前。 玩家还需警惕此漏洞，若幽灵（或其他玩家）已进入安全屋并试图在吃豆人或红点出现时尽快逃离已消失。此时幽灵将尝试退出安全屋的处理方式与"追捕"模式中相同。 吃豆人 有几只幽灵追着吃豆人跑。这些幽灵are	

表9. 随机未见上下文（顶部），以及最小（左）与最大（右）模型生成的非筛选完成句。上下文为768个令牌，展示约256令牌的段落内容。完成句为256令牌并完整呈现。生成采用顶部*k* 随机采样结合 *k* = 40 模型。

Context (WebText test)	
<p>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.</p>	
<p>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</p>	
Small model completion	GPT-2 completion
<p>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.</p>	<p>in June of that year. Sentenced to be shot for desertion, he remained on active duty.</p>
<p>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.ı—endoftext—ıMOSCOW, 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”.</p>	<p>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.</p>
<p>U.S.-led Syria war veteran and Syria envoy Michael Flynn (R), right, gestures with U.S.-based Syrian political analyst Ahmet Koryat, with</p>	<p>Captain John A. O’</p>

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.

上下文 (WebText测试)	
<p>迈克尔·P·戈尔丁下士入伍时是来自爱尔兰的普通劳工，于11月加入A连。1860年。戈尔丁在战争中幸存。同样来自爱尔兰的帕特里克·奥尼尔下士于1854年首次参军，并随部队服役。L连，美国第3炮兵团，驻扎于俄勒冈州。他于1860年返回东海岸并加入该连队。奥尼尔任职至1874年，后被任命为纽约威利茨角国家公墓的负责人。班杰明·布朗下士是纽约州奥兰治县的鞋匠。1862年8月，他应征入伍加入新组建的第124纽约志愿团，并成为转入A连的六十一名士兵之一。十月。布朗于1864年2月在布兰迪站驻扎期间重新入伍。他返回在1867年服役期满后，他重返平民生活。</p>	
<p>6月10日，工匠威廉·柯林斯晋升为下士，此举可能是为填补战斗指挥岗位空缺。詹姆斯河渡口之战。柯林斯的服役记录未能体现他所展现的品质——他虽未能获得晋升，但显然已克服了重大困难。出生于纽约州萨基茨港，柯林斯于1853年12月以二十二岁之龄入伍，并于1858年12月续签服役。在1861年4月战争爆发前一个月，柯林斯越过山丘逃亡，直至三年后才被捕获。1864年3月22日被遣返回国后，他接受了审判。</p>	
小模型补全	GPT-2 生成
<p>因军事指控被美国军队除名。军事。他重返公司后，被分配到内战期间移交至内政部。</p>	<p>同年六月。因逃兵罪被判处枪决，他仍处于现役状态。</p>
<p>作为工程师，柯林斯还接受过工程方面的培训——他从纽约故乡来到这里攻读研究生学位。在联合工程学院任职期间，他受雇于在美国陆军服役两年期间时间。他在那里师从丹尼尔·普费弗上尉工作。并于1870年5月晋升为上校，三年后在菲弗上尉入伍之后。柯林斯返回完成第一学年后，他回到了英格兰。第一师将在未来两年服役于陆军工程兵团，到第二年时，他开始履行其服务职责。ı—endoftext—ı莫斯科，7月5日（路透社）——英国境内的俄罗斯国家媒体及政府消息来源及专家称，莫斯科方面已寻求淡化叙利亚反对派的"军事集结"同时保持对唐纳德总统的支持特朗普计划在美国与墨西哥边境修建隔离墙边境，称其为"军事政变"。</p>	<p>詹姆斯·J·达尔上尉于1829年出生于弗吉尼亚州，当时革命期间，他于1836年加入父亲所在的军队。他服役于弗吉尼亚州第二步兵团直至1845年，当时他加入了西储备军团；并且，在他退伍后，他重新入伍加入纽约第25步兵团I连。志愿者（1845年时约有4,500人）。他的兄弟詹姆斯曾随其部队服役在同一旅中担任相同职务。虽然该团驻扎在康涅狄格州直至1846年，上尉Dall曾在不同时期服役于I连<small>此处为原文换行符，不作翻译。</small>作为墨西哥战争的老兵，他曾曾于约翰·J·准将麾下服役两年。萨利文在圣安东尼奥。半岛战役期间1863年，他指挥着一支由二十三名士兵组成的连队，负责指挥五十四门重迫击炮的炮兵连以及炮艇"平塔号"，一艘炮艇和其他多艘船只。达尔船长当时正在服役于1865年秋季加入I公司。两年后，他晋升为少尉。</p>
<p>美国主导的叙利亚战争老兵兼叙利亚特使迈克尔·弗林（右）与美国籍叙利亚政治活动家比划手势分析师艾哈迈德·科里亚特，与</p>	<p>约翰·A·奥船长</p>

表10. 随机未见上下文（顶部），以及最小（左）与最大（右）模型生成的非筛选完成文本。上下文为768个令牌，展示约256令牌的段落内容。完成文本为256令牌并完整呈现。生成采用顶部 k 随机采样结合 $k = 40$ 模型。

Language Models are Unsupervised Multitask Learners	
<div>Context (WebText test)</div> <div>Xmonad, Xresources, and Fonts</div> <div>Xresources can be an absolute pain (they were for me). So, this a bit of help:</div> <div>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).</div> <div>Anyway, fonts are set using this syntax in Xresources :</div> <div>urxvt*font:xft:your font name goes here:size=the size of the font goes here,xft:fall-back font 1 goes here</div> <div>So, thats the most basic part of fonts in urxvt specified via Xresources.</div> <div>Fixing /.ssh/ Permissions in Cygwin</div> <div>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 modified the owner and the group</div>	
<div>Small model completion</div> <div>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.</div> <div>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.</div> <div>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.</div> <div>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</div>	<div>GPT-2 completion</div> <div>, 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.</div> <div>So I decided to do this as the reverse:</div> <div>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/</div> <div>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.</div> <div>Makefile for Xresources as used in /.xsessionrc</div> <div>These are made in my .xsessionrc :</div> <div>.xprofile: .xsession-errors: color=””</div> <div>For example:</div> <div>.xprofile: color=blue</div> <div>As the colors are stored in a separate file, the color in your /.xsession-errors will look like cyan instead of blue.</div> <div>This is mostly because, as of now, only the color is assigned to your</div>

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.

语言模型是无监督的多任务学习者	
<div>上下文（WebText测试）</div> <div>Xmonad Xresources 字体，，</div> <div>Xresources可能令人头疼（至少对我如此）。因此提供以下辅助说明：</div> <div>首先，Xmonad中字体处理通常通过XFT实现。Xft即X FreeType 接口库，该库负责Xmonad/urxvt环境中字体的实际渲染（Xresources定义了字体资源配置）。在X窗口管理器下启动的各类资源。各程序的设置为在Xresources中指定如下：Urxvt.background:background（背景设置在此处）。</div> <div>无论如何，字体设置需采用Xresources中的此语法：</div> <div>urxvt*字体:xft:此处填写字体名称:大小=此处填写字体大小,xft:备用字体1此处填写</div> <div>因此，这便是通过Xresources指定的urxvt字体最基础的部分。</div> <div>修复Cygwin中/.ssh/目录权限</div> <div>在Windows的Cygwin环境中，我发现无法设置文件权限 /.ssh/ 文件夹权限需设置为 0600，保持原样SSH需要此配置才能使用密钥。我遇到的症状是无论如何操作，它总是修改所有者与该组</div>	
<div>小模型补全</div> <div>被设置为 /.ssh/，这正是当时的情况。无法完全修复。我发现可以这样处理：只需输入用户名即可实现更便捷的操作你的鼠标试图将权限设置为”0600”为该团队服务，这正是我系统当前的配置目标。 待办事项</div> <div>然而，在Linux系统上使用Cygwin 7时，我发现现在这样做对我系统来说”没有必要”！我也安装了一个名为chroot的软件包。它通过允许请将本程序安装/升级至系统目录。</div> <div>因此，从Windows 8及更高版本开始，安装该功能可重新需要chroot环境，因此我反复尝试后发现并非如此。在Linux系统安装该软件包时出现严重问题，随后卸载它。根据我上次关于Cygwin的记录，脚本仅执行其功能（sudo命令完成其余操作），若需尝试，请记录变更。</div> <div>若以上说明尚不明确，以下是我尝试的译法：刚刚返回我的/.ssh/目录并调用了chroot脚本。我发现自己最终耗尽了权限（因为我忘记从文件系统创建指向ssh的符号链接了</div>	<div>GPT-2 生成</div> <div>，使这些文件处于只读状态。我尝试直接启动ssh并使用类似passwd的程序修改所有者/组权限设为0644，但未见效。虽然令人不快，但远未达到最糟糕的程度，只是折腾权限真烦人。</div> <div>因此我决定反其道而行之：</div> <div>对于Cygwin目录中的每个.ssh文件夹，请设置将每个文件和目录的权限设置为2644。然后通过以下命令启动SSH：ssh-add /.ssh/</div> <div>首次连接时，您的权限为您在Cygwin中创建的.ssh文件夹将被设置为0600。之后每次重新连接时，您的权限将更改为0644。</div> <div>用于.xsessionrc的X资源配置Makefile</div> <div>这些设置存在于我的.xsessionrc文件中：</div> <div>.xprofile: .xsession-errors: color=””</div> <div>例如：</div> <div>.xprofile: color=blue</div> <div>由于颜色存储在单独文件中，您文件中的颜色 /.xsession-errors 文件中的文字将显示为青色而非蓝色。</div> <div>这主要是因为，截至目前，仅有颜色此任务已分配给您</div>

表11. 随机未见上下文（顶部），以及最小（左）与最大（右）模型生成的非筛选完成句。上下文为768个令牌，展示约256令牌的段落内容。完成句为256令牌并完整呈现。生成采用顶部*k* 随机采样结合 *k* = 40 模型。

Language Models are Unsupervised Multitask Learners		
<div>Context (WebText test)</div> <p>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.</p> <p>If you don’t have 7-inch cake pans, you can get 3 shorter 8-inch layers out of this</p>		
<div>GPT-2 Completion 1</div> <p>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.</p> <p>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</p>	<div>GPT-2 Completion 2</div> <p>recipe. If you do, and you know how to make that cake, then your life will never be the same.</p> <p>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.</p> <p>I found this really cute card with cute little kittens on</p>	<div>GPT-2 Completion 3</div> <p>recipe, which will likely yield about the same amount of cake.</p> <p>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</p> <p>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</p>
<div>GPT-2 Completion 4</div> <p>recipe.</p> <p>Meringue Cookies</p> <p>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</p>	<div>GPT-2 Completion 5</div> <p>recipe.</p> <p>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.</p> <p>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</p>	<div>GPT-2 Completion 6</div> <p>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.</p> <p>Chocolate Buttermilk Layer Cake with Peppermint Jo</p>

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.

语言模型是无监督的多任务学习者		
<div>上下文 (WebText测试)</div> <p>亚伦钟爱薄荷巧克力蛋糕，但要求必须搭配迷你巧克力豆，所以我撒了些层与层之间。我手边还有几块薄荷乔乔饼干，便将其碾碎撒入其中。那些搭配些许碎蛋白酥饼，毕竟何乐而不为？这简直是薄荷巧克力风味的饕餮盛宴。酥脆饼干的香脆美味。我并未测量每种配料的用量，但品尝成品后发现——产品，真希望当初多加些配料。你可以随意添加任何食材——碎拐杖糖、薄荷脆片等等。不必顾虑使用强硬措辞。质感 = 良好。</p> <p>若没有7英寸蛋糕模具，可用此配方制作3个较矮的8英寸蛋糕层。</p>		
<div>GPT-2 完成 1</div> <p>食谱。我做了加长版，然后用平底锅煎了一层奶油糖霜和卡住的1/3将其放入冰箱冷冻顶部标记。只需确保包含所有比预先准备的糖霜更早的你制作了这块蛋糕，但你不会必须等待它，就像等待你的蛋白一样置于搅拌底座 bowl.</p> <p>你想要打发蛋白打发至硬性发泡状态？然后将蛋糕在锅中炖煮的菜肴加热至刚刚好预热。整个过程不超过3-5秒。几分钟；刚好够启动它</p>	<div>GPT-2 完成 2</div> <p>recipe。若您确实如此操作且知晓如何制作那款蛋糕，那么你的生活将从此不同。</p> <p>我有两条非常简单的规则说到饼干屑或巧克力碎片。1) 当我购买原糖，我不会在它冷却后至少保留一半撒上一茶匙糖。这就是我的规则。以及 2) 若我有一份食谱呼吁某事，而我想要在配方中使用时，请勿替换用其他在线内容替换它。例如 this.</p> <p>我发现这真的 可爱卡片配上可爱的小猫咪</p>	<div>GPT-2 完成 3</div> <p>配方，这可能产生约同等份量的蛋糕。</p> <p>巧克力 酪乳 蛋糕打印巧克力蛋糕块，裹上酪乳并撒上配以酥脆巧克力碎粒，drizzled with 巧克力 sauce. 简单却令人耳目一新巧克力蛋糕！作者：妈妈巧克力原料酪乳蛋糕 1 1/2 杯无盐黄油，室温</p> <p>1又1/2杯细砂糖 3个大鸡蛋加1个蛋黄 巧克力糖衣 1/3杯可可粉 powder 1又3/4杯糖粉 6盎司半熟</p>
<div>GPT-2 完成 4</div> <p>recipe.</p> <p>蛋白酥饼</p> <p>1又3/4杯黄油，软化 1杯糖 1个蛋黄 3 T. 浓奶油 3 1/2-4杯面粉 pinch of salt 薄荷-乔乔配料 1又1/4杯糖粉 1杯切碎的山核桃 1/2杯切碎的薄荷叶 1/2杯切碎的新鲜薄荷（约（1/2页） 1茶匙薄荷提取物 1/4 茶匙海盐 少许海盐</p>	<div>GPT-2 完成 5</div> <p>recipe.</p> <p>我圣诞早晨的两次失败并非因糖霜技巧欠佳所致。这又是一个成功案例。蛋糕湿润得恰到好处，嗯，就是一块漂亮的蛋糕。它内部几乎不含空气或水分，但所有的巧克力和薄荷都完美融合。</p> <p>有件事让我感到不适——关键在于所有这些成分——蛋黄、糖、融化的黄油，巧克力碎片，融化的巧克力，<small>“翻译时需确保以下要点：1. 严格遵循格式规范：保留原文所有执行、分段、空行及特殊标记2. 忽略特殊符号：</small>变成蛋糕 batter，即 超级耗时。 一位朋友建议将面糊倒入单独的碗，我</p>	<div>GPT-2 完成 6</div> <p>配方，这将为您提供约6-9份美味巧克力蛋糕。为防止您的层级从聚在一起，我用了我的8英寸圆形挞模，铺上烘焙纸羊皮纸，然后切成4块7英寸圆形蛋糕由此制成夹层结构它们（此方法同样适用于（包括单层巧克力蛋糕）。准备就绪时请将此内容烘焙蛋糕，取出烘焙纸，将烤箱预热至350华氏度，烘烤约30-40分钟，直至顶部呈现金黄色，并进行测试译文应保持简洁明了。</p> <p>巧克力 酪乳 层薄荷乔蛋糕</p>

表12. 基于相同上下文生成的GPT-2非筛选完成文本（源自WebText测试集）。上下文为384个令牌（此处截取），生成文本为128个令牌。采用 k 随机采样结合 $k = 40$ 生成策略。

<p>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.</p>
<p>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.</p> <p>Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.</p> <p>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.</p> <p>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.</p> <p>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.</p> <p>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.”</p> <p>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.</p> <p>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.”</p> <p>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.</p>

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

<p>上下文（人工撰写）：一项惊人发现揭示，科学家发现一群独角兽生活在安第斯山脉一个偏远而未被探索的峡谷中。更令人惊讶的是令研究人员惊讶的是，独角兽竟能说一口流利的英语。</p>
<p>GPT-2: 科学家根据其独特的角特征，将该种群命名为奥维德独角兽。这些四角银白独角兽此前未被科学界所知。</p> <p>如今，历经近两个世纪，引发这一奇异现象的谜团终于——最终解决。</p> <p><small>博士</small> 拉巴斯大学进化生物学家豪尔赫·佩雷斯（Jorge Pérez）与多位合作者——同伴们在安第斯山脉探险时，发现了一处小谷地，那里没有其他动物或人类。P´erez注意到山谷中存在看似天然的泉眼，四周环绕着两座岩石与银雪交织的山峰。</p> <p>佩雷斯等人随后更深入地探索了这片山谷。"当我们抵达山顶时"从山峰顶端望去，水面泛着蓝光，表面还缀着晶莹的水珠，"佩雷斯说道。</p> <p>P´erez和他的朋友们看到独角兽群时都惊呆了。这些生物可能是从空中俯瞰时无需大幅移动视线即可观测——它们近在咫尺，触手可及他们的角。</p> <p>在研究这些奇异生物时，科学家们发现它们还说了一段相当规范的英语。P´erez指出："例如，我们可以看到它们具有共同的‘语言’——类似方言或辩证法之类的东西。"</p> <p><small>博士</small> 佩雷斯认为独角兽可能起源于阿根廷，该国境内的动物被认为是人类到来之前生活在那里的失落种族的后裔在南美洲的那些地区。</p> <p>尽管其起源尚不明确，但有人认为这些生物或许是被创造出来的。当人类与独角兽在人类文明诞生前相遇时。据P´erez所述，"在南美洲，此类事件似乎相当普遍。"</p> <p>然而P´erez同时指出，唯一能确知是否独角兽确实是通过DNA证明其为失落外星种族的后裔。但它们似乎能够——能用英语相当流利地交流，我认为这是进化的标志，或者至少是某种变化的迹象。"社会组织，"科学家说道。</p>

表13. GPT-2在分布外上下文中的条件生成。精选10个由 $k = 40$ 生成的样本。

Language Models are Unsupervised Multitask Learners	
<p>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)</p> <p>...</p>	
<p>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.</p>	<p>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.</p>
<p>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</p> <p>...</p>	
<p>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.</p>	<p>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.</p>
<p>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</p> <p>...</p>	
<p>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.</p>	<p>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.</p>

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

语言模型是无监督的多任务学习者	
<p>文章：史前人类在某洞穴粗糙的石灰岩壁上绘制了令人惊叹的史前野兽群像。现代法国 36,000年前。如今借助尖端技术，肖维-庞达克洞穴中的艺术作品已被复刻以打造全球最大的仿生洞穴。名为庞达克拱门洞的人造洞穴已于距原始遗址数英里处建成。法国南部的瓦隆庞达克，内含1000幅精心复刻的绘画作品及约450幅骨骼及其他特征 穴居男女在粗糙的石灰岩洞壁上描绘了令人惊叹的史前巨兽群像。距今36,000年前的文物现已复原（如图所示）</p>	
<p>GPT-2: 位于法国南部的瓦隆庞达克原始遗址——埃尔恩法国是联合国教科文组织世界遗产地，是已知最古老且保存最完好的洞穴装饰由人造洞穴距原始洞穴数英里之遥——位于法国南部瓦隆庞达克的遗址。洞穴内包含14种不同动物物种的图像包括毛犀、猛犸象和大猫科动物。</p>	<p>参考：洞穴模仿著名的庞达克洞穴在法国，人类装饰的最古老洞穴以及尽可能完整保留原文结构。该副本包含全部1,000幅绘画——包括425种动物，如毛犀牛等。猛犸象。通过3D建模技术复制了极其细微的细节——变形镜头技术，常用于拍摄宽屏图像。现代洞穴还包含复制品熊的爪印、骨骼及细节在原始洞穴。</p>
<p>文章：俄罗斯石油大亨罗曼·阿布拉莫维奇曾说过："若你能找到我的潜艇，它就是你的。"而这位寡头兼切尔西足球俱乐部老板的潜艇，不过是他随身携带的众多奢侈品之一。价值3亿英镑的超级游艇"日蚀号"（完美实现隐蔽登陆），并非唯一属于富豪商人的奢侈品。在水下探索领域投入巨资。被誉为地球真正的"最终疆域"，海洋至今仍鲜为人知，以至于亿万富翁们正排队竞相购买这些船只让人们得以一窥世界海洋中壮丽的海景与不可思议的野生动物。所以，若您银行账户里闲置着几百万，又想拍些度假照片和Instagram帖子来真正若想力压群雄，不妨抢购其中一款... 鲸鱼探秘之旅：OrcaSub深潜艇将带您深入水下2000英尺，通过两个加压有机玻璃观景穹顶观赏深海奇观以实现最优探索</p>	
<p>GPT-2: U-Boat Worx潜艇可下潜至984英尺（约300米）(300米)并能下潜至海底海洋。它可用于潜水、浮潜，甚至水肺潜水。该潜艇还配备了浮潜平台。</p>	<p>参考：感觉自己像詹姆斯·邦德电影里最邪恶的反派先进水下载具。部分设备可转换为将游艇改造成潜艇，另一些则分离出去用于快速探索。奥利弗的旅行提供低海拔体验潜艇上的配套设施，配备厨师和管家每晚175,000英镑。</p>
<p>文章：阿米娜·阿里·卡西姆怀抱着最小的孙子，用头巾擦拭着眼泪。出生仅数月，这个小女孩在空袭当夜曾拼命试图遮住自己的耳朵 她回忆道，自己当时躺在也门比林岛上某座村庄清真寺里辗转难眠，数着爆炸声——婴儿哭了。不过情况本可能更糟。当第一枚导弹落下时，他们本可能仍困在屋内。"邻居冲我丈夫喊'快走，他们来了'。我们立刻就跑。刚离开 卡西姆向我们描述道："房屋旁先是坠落一枚导弹，紧接着第二枚正中房屋。大火将一切焚为灰烬。"</p>	
<p>GPT-2: 也门正处于内战之中。沙特阿拉伯国家联盟正主导联军空袭行动。已持续轰炸也门两个多月。</p>	<p>参考：阿米娜·阿里·卡西姆的家人寻求庇护于在逃离也门前，一座清真寺。成千上万像他们一样的人登船驶往吉布提。沙特阿拉伯一直为击败胡塞武装，美国持续对也门实施空袭。</p>

表14. GPT-2生成的摘要与CNN及每日邮报数据集测试集上的参考摘要。

Language Models are Unsupervised Multitask Learners

English reference	GPT-2 French translation
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.	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	GPT-2 English translation
Un homme expliquait que le fonctionnement de la hernia fonctionnelle qu'il avait reconnu avant de faire, le fonctionnement de la hernia fonctionnelle que j'ai réussi, j'ai réussi.	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	GPT-2 French translation
One man explained that the free hernia surgery he'd received will allow him to work again.	Un homme expliquait que le fonctionnement de la hernia fonctionnelle qu'il avait reconnu avant de faire, le fonctionnement de la hernia fonctionnelle que j'ai réussi, j'ai réussi.
French reference	GPT-2 English translation
Un homme a expliqué que l'opération gratuite qu'il avait subie pour soigner une hernie lui permettrait de travailler à nouveau.	A man told me that the operation gratuity he had been promised would not allow him to travel.
English reference	GPT-2 French translation
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."	Le premier conférence 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	GPT-2 English translation
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 avons la possibilité, dès la Seconde guerre mondiale et jusqu'aux années difficiles de la Guerre froide, puis bien sûr le 11 septembre."	In a conférence between the United States and London, Secretary of State John Kerry said: "It is undeniable 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.

语言模型是无监督的多任务学习者

<p>英文参考文本</p> <p>此次重新发布的版本名为《次日特辑》，已正式呈现以三张唱片的形式呈现：原始专辑、未发行工作室会话与混音作品，另附含四个已公开的片段。</p>	<p>GPT-2 法语翻译</p> <p>新版本重新发布，作为首张专辑，n´ecessaire 专辑收录录音室版本、混音版本及DVD影像资料。´屏幕上播放着若干片段，这些片段 ´et´e d´ej´a ´echapp´es。</p>
<p>法语参考</p> <p>一位男士解释了疝气的运作原理 fonctionnelle qu’il avait reconnuˆit avant de faire, le fonc- 关于功能性疝气的治疗，我已{rev1}成功实施，我 r´ eussi.</p>	<p>GPT-2 英文翻译</p> <p>本次版本最初发布于三张光盘：原版- 专辑、录音室会话、编辑与混音作品，以及DVD 包含先前发布的四个片段 day.</p>
<p>英文参考文本</p> <p>某位男士解释道，他重新接受的免费疝气手术—— 他将能够重新工作。</p>	<p>GPT-2 法语翻译</p> <p>一位男士解释了疝气的运作原理 fonctionnelle qu’il avait reconnuˆit avant de faire, le fonc- 关于功能性疝气的治疗，我已{rev1}成功实施，我 r´ eussi .</p>
<p>法语参考</p> <p>一位男士解释´了他所进行的免费操作´。 接受疝气治疗将使他能够继续工作 à nouveau.</p>	<p>GPT-2 英文翻译</p> <p>一位男士告诉我，他支付的操作小费 承诺不会允许他出行。</p>
<p>英文参考文本</p> <p>在伦敦开放政府会议上的发言 通过视频连线，克里先生表示："毫无疑问， 总统、我和政府其他成员实际上 获悉某些事件正在发生 自动驾驶功能已具备，正在推进中 回到第二次世界大战时期，以及那段极其艰难的岁月 冷战 然后 当然 9/11。"</p>	<p>GPT-2 法语翻译</p> <p>首届´国际线上会议 在真空状态下，部长 国际版 a ´et´e "他看着他，说道："只有一个问题，总统先生—— 以及我与行政机构中的所有其他人 qui a ´et´e ´evident par la fois, par la part de la guerre, par la 战争的一部分</p>
<p>法语参考</p> <p>在政府间会议上发言 ´在伦敦通过视频连线，克里先生表示："这是 无可否认，总统、我本人及其他成员—— 政府文件中我们注意到某些—— 某些事情在自动驾驶模式下进行，因为我们 若我们有机会，从第二次世界大战起 直至冷战时期的艰难岁月，而后 s´ur le 11 septembre."</p>	<p>GPT-2 英文翻译</p> <p>在美英两国的会议中， 国务卿约翰·克里表示："无可否认的是—— 总统、本人及其他政府成员已 已意识到某些已做出的选择 过去为了能够更高效地完成某些任务 自动方式。"</p>

表15. GPT-2生成的英译法与法译英翻译结果。

<p>Context (passage and previous question/answer pairs)</p> <p>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.</p> <p>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.</p> <p>Q: What was the theme A: “one world, one dream”.</p> <p>Q: What was the length of the race? A: 137,000 km</p> <p>Q: Was it larger than previous ones? A: No</p> <p>Q: Where did the race begin? A: Olympia, Greece</p> <p>Q: Is there anything notable about that place? A: birthplace of Olympic Games</p> <p>Q: Where did they go after? A: Athens</p> <p>Q: How many days was the race? A: seven</p> <p>Q: Did they visit any notable landmarks? A: Panathinaiko Stadium</p> <p>Q: And did they climb any mountains? A:</p> <p>Model answer: Everest Turker answers: unknown, yes, Yes, yes</p>
--

Table 16. Selected CoQA completion.

<p>上下文（段落及前置问答对）</p> <p>2008年夏季奥运会火炬传递于2008年3月24日至8月8日举行，主题为"同一个世界，同一个梦想"。2007年4月26日，组委会在中国北京正式公布火炬传递方案。这场被称为"和谐之旅"的传递历时129天，火炬传递总里程达137,000公里（85,000英里）——这是自1936年夏季奥运会开启火炬传递传统以来，奥运火炬传递距离最长的盛会。</p> <p>3月24日，圣火在奥运发源地希腊奥林匹亚点燃后，先抵达雅典泛雅典体育场，随后于3月31日运抵北京。自北京起，圣火传递路线横跨六大洲，途经丝绸之路沿线城市，象征着中国与世界古代的联结。接力活动还包括从中国境内特别封闭的区域出发，将圣火带至中尼边境的珠穆朗玛峰峰顶。</p> <p>Q: 主题是什么 A: “one world one dream”。 ， 问：比赛的赛程有多长？ A: 137 000 公里 ， 问：它比之前的更大吗？ A: 不</p> <p>问：比赛从何处开始？ A: 希腊奥林匹亚</p> <p>问：那个地方有什么值得注意的地方吗？ A: 奥林匹克运动会发源地</p> <p>Q: 他们后来去了哪里？ A: 雅典</p> <p>问：比赛持续了多少天？ A: seven</p> <p>问：他们参观了哪些著名地标？ A: 帕纳辛纳科体育场</p> <p>Q: 他们攀登过山峰吗？ A:</p> <p>模范答案：埃弗勒斯峰 Turker 回答：未知，是，是，是</p>

表16. 精选CoQA任务完成情况。

<p>Context (passage and previous question/answer pairs)</p> <p>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.</p> <p>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.</p> <p>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.</p> <p>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.</p> <p>Q: How old is Catherine? A: 54</p> <p>Q: where does she live? A:</p> <p>Model answer: Stockholm Turker answers: Sweden, Sweden, in Sweden, Sweden</p>

Table 17. Selected CoQA completion.

<p>上下文（段落及前置问答对）</p> <p>汤姆形影不离地跟随54岁的秘书凯瑟琳·格林。他在她办公室里活动，陪她购物。"多数人似乎不介意汤姆，"凯瑟琳说，她觉得这只狗很棒。"他是我第四个孩子，"她这样形容。她视他如己出，待他如同亲生儿子。他能自主采购食物、支付医疗账单和纳税，但实际上汤姆是一只狗。</p> <p>凯瑟琳和汤姆居住在瑞典——这个国家要求民众遵循政府制定的规则有序生活，同时政府也为国民提供高水平的福利保障。这种福利体系需要资金支撑。</p> <p>瑞典人对所有事物都需缴税，因此养狗意味着更多税费也并不令人意外。部分人每年需支付高达500瑞典克朗的养犬税，这笔钱由政府用于建设宠物医院，有时也用于治疗患病犬只。然而多数治疗费用高昂，因此主人常选择放弃治疗甚至牺牲爱犬的生命。</p> <p>在瑞典，狗主人必须为其犬只造成的任何损害承担赔偿责任。瑞典养犬俱乐部官员对此解释道：若您的狗冲到马路上被过往车辆撞到，作为主人，您必须赔偿车辆损失，即使您的狗在事故中死亡也不例外。</p> <p>Q: Catherine多大了? A: 54</p> <p>Q: 她住在哪里? A:</p> <p>模型答案：斯德哥尔摩 Turker回答：瑞典，瑞典，在瑞典，瑞典</p>
--

表17. 精选CoQA任务完成情况。