# Language Models are Few-Shot Learners

Tom B. Brown*， Benjamin Mann*， Nick Ryder*， Melanie Subbiah*，
Jared Kaplan， Prafulla Dhariwal， Arvind Neelakantan， Pranav Shyam， Girish Sastry，
Amanda Askell， Sandhini Agarwal， Ariel Herbert-Voss， Gretchen Krueger， Tom Henighan，
Rewon Child， Aditya Ramesh， Daniel M. Ziegler， Jeffrey Wu， Clemens Winter，
Christopher Hesse， Mark Chen， Eric Sigler， Mateusz Litwin， Scott Gray，
Benjamin Chess， Jack Clark， Christopher Berner，
Sam McCandlish， Alec Radford， Ilya Sutskever， Dario Amodei
OpenAI

Abstarct

Recent work has demonstrated substantial gains on many NLP tasks and benchmarks by pre-training on a large corpus of text followed by fine-tuning on a specific task. While typically task-agnostic in architecture, this method still requires task-specific fine-tuning datasets of thousands or tens of thousands of examples. By contrast, humans can generally perform a new language task from only a few examples or from simple instructions – something which current NLP systems still largely struggle to do. Here we show that scaling up language models greatly improves task-agnostic,few-shot performance, sometimes even reaching competitiveness with prior state-of-the-art fine-tuning approaches. Specifically, we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model, and test its performance in the few-shot setting. For all tasks, GPT-3 is applied without any gradient updates or fine-tuning,with tasks and few-shot demonstrations specified purely via text interaction with the model. GPT-3 achieves strong performance on many NLP datasets, including translation, question-answering, and cloze tasks, as well as several tasks that require on-the-fly reasoning or domain adaptation, such as unscrambling words, using a novel word in a sentence, or performing 3-digit arithmetic. At the same time, we also identify some datasets where GPT-3’ s few-shot learning still struggles, as well as some datasets where GPT-3 faces methodological issues related to training on large web corpora. Finally,we find that GPT-3 can generate samples of news articles which human evaluators have difficultydistinguishing from articles written by humans. We discuss broader societal impacts of this finding and of GPT-3 in general.

近期的工作表明，通过在大语料上进行预训练，然后微调，可以很多NLP任务和基线上有大幅提升。不过这些研究虽说在架构上与任务无关，但是仍然需要成千上万的数据来微调。仅仅通过一少部分样例或简单的提示，人类就可以完成一个新的语言任务，但是，当前的NLP系统其实很难做到。在这里，我们证明了，提高语言模型的规模可以大幅提升任务无关的few_shot表现，有时甚至能达到之前微调模型的最佳成绩。我们训练了GPT-3，一个有1750亿个参数的自回归语言模型，它的大小为之前非稀疏模型的10倍之多，同时测试了它在few_shot设置下的表现。对于所有任务，GPT-3都是在没有

更新权重或微调的情况下使用的，仅仅是通过文本和模型沟通。结果，GPT-3在很多NLP数据上取得了很好的表现，包括翻译、问答和完形填空，还有一些即时推理或领域适应的场景，比如拼词、句子中使用新词、3位数数学运算。同时，我们还确定了一些数据，GPT-3的few_shot仍然难以应对，还有一些数据，GPT-3是需要面对类似大的web语料的一些问题。最后，我们发现GPT-3可以生产新闻文章，这些文章甚至人类评估者都很难辨别是否出自人类。我们还讨论了这一发现和GPT-3的更大的社会影响。

Contents

目录

1、引言

Recent years have featured a trend towards pre-trained language representations in NLP systems, applied in increasingly flexible and task-agnostic ways for downstream transfer. First, single-layer representations were learned using wordvectors [MCCD13, PSM14] and fed to task-specific architectures, then RNNs with multiple layers of representations and contextual state were used to form stronger representations [DL15, MBXS17, PNZtY18] (though still applied to task-specific architectures), and more recently pre-trained recurrent or transformer language models [VSP+17] have been directly fine-tuned, entirely removing the need for task-specific architectures [RNSS18, DCLT18, HR18].

近些年，NLP领域出现了一个趋势，即先进行模型预训练，然后灵活地运用在下游任务上。起先，是有wordvectors [MCCD13, PSM14]这样的单层的表示方法，这种表示方法接着送进特殊任务的网络架构，后来，出现了多层表示和上下文状态来构造更强的表示[DL15, MBXS17, PNZtY18] ，即RNN（然后仍然需要特殊网络架构），最近，预训练的循环或者transformer语言模型 [VSP+17] 可以直接微调，完全去除了特殊网络架构[RNSS18, DCLT18, HR18]。

This last paradigm has led to substantial progress on many challenging NLP tasks such as reading comprehension,question answering, textual entailment, and many others, and has continued to advance based on new architectures and algorithms [RSR+19, LOG+19, YDY+19, LCG+19]. However, a major limitation to this approach is that while the architecture is task-agnostic, there is still a need for task-specific datasets and task-specific fine-tuning: to achieve strong performance on a desired task typically requires fine-tuning on a dataset of thousands to hundreds of thousands of examples specific to that task. Removing this limitation would be desirable, for several reasons.

最后一种范式已经在许有挑战的NLP任务上取得了大幅进步，比如阅读理解、问答、文本蕴含以及其它，并且在新的架构和算法[RSR+19, LOG+19, YDY+19, LCG+19]上也有提升。然而，这种方法有一个主要的限制，就是虽然后架构上来说是任务无关的，但是仍然有必要对任务相关的数据进行微调：要想在特殊任务上微调以取得好的效果，就需要成千上万的样本。如果能去掉这个限制，将会是令人激动地，这有几个原因。

First, from a practical perspective, the need for a large dataset of labeled examples for every new task limits the applicability of language models. There exists a very wide range of possible useful language tasks, encompassing anything from correcting grammar, to generating examples of an abstract concept, to critiquing a short story. For many of these tasks it is difficult to collect a large supervised training dataset, especially when the process must be repeated for every new task.

首先，从实践的角度来说，对每一个新任务来说，大量带标注的数据限制了语言模型的应用。有很多有用的语言任务，包括语法纠错，生成抽象概念，评论短篇小说 。有很多这样的任务是很难收集监督训练数据集的，尤其是对第一个新任务都要再收集一遍。

Second, the potential to exploit spurious correlations in training data fundamentally grows with the expressiveness of the model and the narrowness of the training distribution. This can create problems for the pre-training plus fine-tuning paradigm, where models are designed to be large to absorb information during pre-training, but are then fine-tuned on very narrow task distributions. For instance [HLW+20] observe that larger models do not necessarily generalize better out-of-distribution. There is evidence that suggests that the generalization achieved under this paradigm can be poor because the model is overly specific to the training distribution and does not generalize well outside it[YdC+19, MPL19]. Thus, the performance of fine-tuned models on specific benchmarks, even when it is nominally at human-level, may exaggerate actual performance on the underlying task [GSL+18, NK19].
其次，随着预训练模型的拟和能力和训练数据的分布狭隘程度提升，学习到虚假相关性的可能性会提升。这会产生一个问题，就是预训练模型是很大，但是微调是在一个很狭窄的任务上。比如[HLW+20]发现较大地模型在分布外不一定表现良好。有证据表明在这种范式下泛化能力很难提升，因为这个模型过度拟和了训练数据的分布，在此之外很难有好的泛化能力[YdC+19, MPL19]。因此，在基准任务上的微调表现，虽然号称可以接近人类水平，也可能是夸大了它的实际性能 [GSL+18, NK19]。
Third, humans do not require large supervised datasets to learn most language tasks – a brief directive in natural language (e.g. “please tell me if this sentence describes something happy or something sad”) or at most a tiny number of demonstrations (e.g. “here are two examples of people acting brave; please give a third example of bravery”) is often sufficient to enable a human to perform a new task to at least a reasonable degree of competence. Aside from pointing to a conceptual limitation in our current NLP techniques, this adaptability has practical advantages – it allows humans to seamlessly mix together or switch between many tasks and skills, for example performing addition during a lengthy dialogue. To be broadly useful, we would someday like our NLP systems to have this same fluidity and generality.
第三，人类不需要大量的监督数据来学习语言任务，比如简短的指令（例如：请告诉我这句话是快乐的还是悲伤的）或者最多只需要少量的例子（例如：这里有两个人们表现勇敢的例子，请给出第三个关于表现勇敢的例子）就足够人类去完成一个新任务，至少能达到合理的水平。除了指出当前语言模型的概念限制外，这种适应性还有实践的优势，它允许人类将许多任务和技能无缝连接和转换，例如在长文本对话中进行加法计算。更广泛地来说，我们希望有一天我们的NLP系统具备这种流动性和通用性。
One potential route towards addressing these issues is meta-learning – which in the context of language models means the model develops a broad set of skills and pattern recognition abilities at training time, and then uses those abilities at inference time to rapidly adapt to or recognize the desired task (illustrated in Figure 1.1). Recent work [RWC+19]attempts to do this via what we call “in-context learning”, using the text input of a pretrained language model as a form of task specification: the model is conditioned on a natural language

instruction and/or a few demonstrations of the task and is then expected to complete further instances of the task simply by predicting what comes next.
解决这些问题的一个有潜力的技术路线就是元学习，这在语言模型的背景下意味着，模型在训练时学习到了一系列的技能和模式识别能力，然后应用这些技能到推理上，以快速地适应或识别所需任务（如图1.1所示）。近期的工作[RWC+19]试图通过我们称之为上下文学习来实现这个目的，使用预训练模型的输入作为特定任务的一种形式：模型基于自然语言指令和/或一系列任务案例为条件，然后期望预测接下来会发生什么来完成任务的输出。
While it has shown some initial promise, this approach still achieves results far inferior to fine-tuning – for example[RWC+19] achieves only 4% on Natural Questions, and even its 55 F1 CoQa result is now more than 35 points behind the state of the art. Meta-learning clearly requires substantial improvement in order to be viable as a practical method of solving language tasks.
虽然它有一些潜力，但是这个方法依旧和微调相差很多，比如[RWC+19]在问答上仅有4%的结果，在CoQa上F1达到55，和最先进的水平差了35个点。元学习仍有大幅提升效果才能解决实际的语言任务。
Another recent trend in language modeling may offer a way forward. In recent years the capacity of transformer language models has increased substantially, from 100 million parameters [RNSS18], to 300 million parameters[DCLT18], to 1.5 billion parameters [RWC+19], to 8 billion parameters [SPP+19], 11 billion parameters [RSR+19],and finally 17 billion parameters [Tur20]. Each increase has brought improvements in text synthesis and/or downstream NLP tasks, and there is evidence suggesting that log loss, which correlates well with many downstream tasks, follows a smooth trend of improvement with scale [KMH+20]. Since in-context learning involves absorbing many skills and tasks with in the parameters of the model, it is plausible that in-context learning abilities might show similarly strong gains with scale.
语言模型还有另一个比较不错的方向。近些年，transformer语言模型的规模在大幅增加，从100M参数 [RNSS18]，到300M参数[DCLT18]，1.5B参数[SPP+19]，11B参数 [RSR+19]，以及最后17B参数[Tur20]。每一次的规模增加都带来了文本合成以及下游NLP任务的效果提升，有证据表明和下游任务相关的对数损失，随着规模的增加呈现平稳的下降[KMH+20]。由于上下文学习涉及从模型参数中学习技能和任务，因此上下文学习的能力和规模强相关。
In this paper, we test this hypothesis by training a 175 billion parameter autoregressive language model, which we callGPT-3, and measuring its in-context learning abilities. Specifically, we evaluate GPT-3 on over two dozen NLP datasets,as well as several novel tasks designed to test rapid adaptation to tasks unlikely to be directly contained in the training set. For each task, we evaluate GPT-3 under 3 conditions: (a) “few-shot learning”, or in-context learning where we allow as many demonstrations as will fit into the model’s context window (typically 10 to 100), (b) “one-shot learning”,where we allow only one demonstration, and

(c) “zero-shot” learning, where no demonstrations are allowed and only an instruction in natural language is given to the model. GPT-3 could also in principle be evaluated in the traditional fine-tuning setting, but we leave this to future work.
在本文，我们验证了这个假设，训练了个1750亿参数的自回归语言模型，称之为GPT-3，而且评估了它的上下文学习能力。我们在20多个NLP数据集上进行了评估，还在一些新任务进行了评估，这些新任务是训练集以外用来测试快速适应性的。对每个任务，我们依据三个条件进行评估：（a）少样本学习或上下文学习，我们将尽可能多的样本输入模型的内容窗口（10到100个不等），（b）单样本学习，（c）零样本学习，仅有一个提示给模型。GPT-3模型在理论上也可以用于传统的微调，但是我们将这个工作留给未来。
Figure 1.2 illustrates the conditions we study, and shows few-shot learning of a simple task requiring the model to remove extraneous symbols from a word. Model performance improves with the addition of a natural language task description, and with the number of examples in the model’s context, K. Few-shot learning also improves dramatically with model size. Though the results in this case are particularly striking, the general trends with both model size and number of examples in-context hold for most tasks we study. We emphasize that these “learning” curves involve no gradient updates or fine-tuning, just increasing numbers of demonstrations given as conditioning.
图1.2展示了我们研究的条件，和一个简单任务的few-shot学习，这个任务是从一个单词中移除多余的符号。模型表现随着任务描述及上下文数量K的增加而提升。Few-shot学习同时也会随着模型大小的增加而提升效果。尽管这个结果在这个例子中特别夺目，但是我们研究的大多数项目都是这个规律。我们强调这些学习曲线不涉及梯度更新或微调，仅仅提升示例的数量即可。

图1.2 更大的模型更有效地利用上下文信息

Broadly, on NLP tasks GPT-3 achieves promising results in the zero-shot and one-shot settings, and in the the few-shot setting is sometimes competitive with or even occasionally surpasses state-of-the-art (despite state-of-the-art being held by fine-tuned models). For example, GPT-3 achieves 81.5 F1 on CoQA in the zero-shot setting, 84.0 F1 on CoQA in the one-shot setting, 85.0 F1 in the few-shot setting. Similarly, GPT-3 achieves 64.3% accuracy on TriviaQA in the zero-shot setting, 68.0% in the one-shot setting, and 71.2% in the few-shot setting, the last of which is state-of-the-art relative to fine-tuned models operating in the same closed-book setting.
更广泛地说，GPT-3在zero-shot和one-shot设置下取得了有希望的结果，在few-shot设置下有时也很有竞争力，甚至有时会超越现有的最好水平（尽管现有的最好水平是微调出来的）。例如，GPT-3在CoQA上，zero-shot取了85.1的F1值，在one-shot上取季84.0的F1值，在few-shot上取得了85.0的F1值。同样地，在TrivialQA任务上，在zero-shot上取得了64.3%的准确率，在one-shot上取得了68.0%的准确率，在few-shot上取得了71.2%的准确率，其中few-shot的结果是在闭书设置情况下超越了微调的结果。

GPT-3 also displays one-shot and few-shot proficiency at tasks designed to test rapid adaption or on-the-fly reasoning,which include unscrambling words, performing arithmetic, and using novel words in a sentence after seeing them defined only once. We also show that in the few-shot setting, GPT-3 can generate synthetic news articles which human evaluators have difficulty distinguishing from human-generated articles.

GPT-3同样展示了在快速适应和即兴推理方面的one-shot和few-shot能力，包括拼字、算法和仅看到一次定义后使用新词造句。我们同样展示了在few-shot设置下，GPT-3可以生成合成新闻，人类评估员都很难将合成新闻和人类写的新闻区分开来。

At the same time, we also find some tasks on which few-shot performance struggles, even at the scale of GPT-3. This includes natural language inference tasks like the ANLI dataset, and some reading comprehension datasets like RACE or QuAC. By presenting a broad characterization of GPT-3’s strengths and weaknesses, including these limitations, we hope to stimulate study of few-shot learning in language models and draw attention to where progress is most needed.

同时，我们发现了一些任务在few-shot上的表现欠佳，即使在GPT-3上也如此。这包括像ANLI这样的推理任务，也包括像RACE和QuAC这样的阅读理解任务。通过对GPT-3的优势和劣势的广泛研究，我们希望引起业界对few-shot学习的兴趣，并关注进展。

A heuristic sense of the overall results can be seen in Figure 1.3, which aggregates the various tasks (though it should not be seen as a rigorous or meaningful benchmark in itself).We also undertake a systematic study of “data contamination” – a growing problem when training high capacity models on datasets such as Common Crawl, which can potentially include content from test datasets simply because such content often exists on the web. In this paper we develop systematic tools to measure data contamination and quantify its distorting effects. Although we find that data contamination has a minimal effect on GPT-3’s performance on most datasets, we do identify a few datasets where it could be inflating results, and we either do not report results on these datasets or we note them with an asterisk, depending on the severity.

图1.3展示了一个具有启示意义的结论，它集合了不同的任务（尽管它不应该被视为严格的或有意义的基本准）。我们还系统研究了“数据污染”，这在比如爬虫数据上训练高容量模型时，是一个日益凸显的问题，因为爬虫数据是来自网络，训练数据会出现在测试数据中。在本文，我们建立了系统的工具来衡量数据污染，并平衡其扭曲效应。尽管我们发现数据污染在GPT3的表现上影响很小，我们还是发现了一些数据集结果有些夸大了，对于这些数据，我们要么不报告其结果，要么将其标星，这取决于它的严重程度。

图1.3 42个以准确率为度量的基本综合测试表现

In addition to all the above, we also train a series of smaller models (ranging from 125 million parameters to 13 billion parameters) in order to compare their performance to GPT-3 in the

zero, one and few-shot settings. Broadly, for most tasks we find relatively smooth scaling with model capacity in all three settings; one notable pattern is that the gap between zero-, one-, and few-shot performance often grows with model capacity, perhaps suggesting that larger models are more proficient meta-learners.
除了上面的这些，我们还训练了一系列的小模型（从1.25亿到130亿的参数）用来和GPT3在zero、one、few-shot上进行对比。一般来说，对于大部分任务，在这三种设置下，我们发现随着模型容量变化效果稳定增长。一个值得注意的一点是，随着模型容量的加大zero、one、few-shot的区别越来越大，也许这表明了更大的模型是更有效的元学习者。
Finally, given the broad spectrum of capabilities displayed by GPT-3, we discuss concerns about bias, fairness, and broader societal impacts, and attempt a preliminary analysis of GPT-3’s characteristics in this regard.
最后，鉴于GPT-3的广泛能力，我们讨论了偏见、公平以及广泛的社会影响，而且尝试对这些特征进行了GPT-3特性的初步分析。
The remainder of this paper is organized as follows. In Section 2, we describe our approach and methods for training GPT-3 and evaluating it. Section 3 presents results on the full range of tasks in the zero-, one- and few-shot settings.Section 4 addresses questions of data contamination (train-test overlap). Section 5 discusses limitations of GPT-3.Section 6 discusses broader impacts. Section 7 reviews related work and Section 8 concludes.
本篇论文的剩余部分结构如下。在第二节，我们描述了我们的训练和评估GPT-3的方法。第三节展示了所有任务在zero-、one- and few-shot设置下的评估结果。第四节讨论了数据污染的问题。第五节讨论了GPT-3的限制。第六节讨 论了其广泛的影响。第七节回顾了相关的工作，第八节进行了总结。
2 方案
Our basic pre-training approach, including model, data, and training, is similar to the process described in [RWC+19],with relatively straightforward scaling up of the model size, dataset size and diversity, and length of training. Our use of in-context learning is also similar to [RWC+19], but in this work we systematically explore different settings for learning within the context. Therefore, we start this section by explicitly defining and contrasting the different settings that we will be evaluating GPT-3 on or could in principle evaluate GPT-3 on. These settings can be seen as lying on a spectrum of how much task-specific data they tend to rely on. Specifically, we can identify at least four points on this spectrum (see Figure 2.1 for an illustration):
我们的基本预训练方案，包括模型、数据和训练和之前的论文[RWC+19]很相似，只是简单地增大了模型大小、数据的大小和多样性，以及训练时长。我们的上下文学习和[RWC+19]也很相似，只是在我们这一次工作中，针对上下文学习，我们系统地探索了不同的设置。因此，本小节从定义和对比不同的设置开始，这些设置是评估GPT-3时需要的。These settings can be seen as lying on a spectrum of how much task-specific data they tend to rely on. Specifically, we can identify at least four points on this spectrum (see Figure 2.1 for an illustration):（待翻译）

Fine-Tuning (FT) has been the most common approach in recent years, and involves updating the weights of a pre-trained model by training on a supervised dataset specific to the desired task. Typically thousands to hundreds of thousands of labeled examples are used. The main advantage of fine-tuning is strong performance on many benchmarks. The main disadvantages are the need for a new large dataset for every task, the potential for poor generalization out-of-distribution [MPL19], and the potential to exploit spurious features of the training data [GSL+18, NK19], potentially resulting in an unfair comparison with human performance. In this work we do not fine-tune GPT-3 because our focus is on task-agnostic performance, but GPT-3 can be fine-tuned in principle and this is a promising direction for future work.

**微调（FT）**是近些年来最常用的方法之一，它通过特定的任务的数据来更新预训练模型的权重。通常使用数千到数十万打标数据。微调的主要优势是在许多基准测试中的强大的性能。主要缺点是需要为每个任务准备一个新的大型数据集，这可能会导致分布外泛化能力差，还可能会学习到训练数据的虚假特征，从而导致与人类的评估的不公平差异。在本论文中，我们不微调GPT-3，因为我们的重点是关注与任务无关的表现，但是理论上GPT-3是可以被微调的，这是一个有潜力的方向。

Few-Shot (FS) is the term we will use in this work to refer to the setting where the model is given a few demonstrations of the task at inference time as conditioning [RWC+19], but no weight updates are allowed.As shown in Figure 2.1, for a typical dataset an example has a context and a desired completion (for example an English sentence and the French translation), and few-shot works by giving K examples of context and completion, and then one final example of context, with the model expected to provide the completion. We typically set K in the range of 10 to 100 as this is how many examples can fit in the model’s context window(nctx = 2048). The main advantages of few-shot are a major reduction in the need for task-specific data and reduced potential to learn an overly narrow distribution from a large but narrow fine-tuning dataset. The main disadvantage is that results from this method have so far been much worse than state-of-the-art fine-tuned models. Also, a small amount of task specific data is still required. As indicated by the name, few-shot learning as described here for language models is related to few-shot learning as used in other contexts in ML [HYC01, VBL+16] – both involve learning based on a broad distribution of tasks (in this case implicit in the pre-training data) and then rapidly adapting to a new task.

Few-Shot（FS）是我们在本研究中使用的设置，在推理时给模型一些样本，但是权重并不更新。如图2.1所示，一个样例包含一个文本及一个期望的输出（比如一个英语的句子和一个法语的翻译），few-shot工作原理为，给出K个包含文本和输出的样例，还有最后一个样例的文本，期望模型提供其输出。我们一般将k设置为10到100，因为这是模型上下文所能容纳的示例数量。few-shot的主要优势是减少了特定任务数据的需要，同时降低了从大量有偏的数据中学习到过拟和模型的可能性。主要缺点是这个方法的结果目前比最好的微调模型还要差。同时，少量的任务相关的数据还是需要的。正如其名，这里语言模型的few-shot学习和机器学习中其它基于上下文的few-shot [HYC01, VBL+16]学习一样，

都涉及基于一个广泛分布的任务进行学习（在这个案例中是预训练数据中的隐式学习），然后快速适应新任务。

图2.1 Zero-shot, one-shot and few-shot和微调的对比

One-Shot (1S) is the same as few-shot except that only one demonstration is allowed, in addition to a natural language description of the task, as shown in Figure 1. The reason to distinguish one-shot from few-shot and zero-shot (below) is that it most closely matches the way in which some tasks are communicated to humans.For example, when asking humans to generate a dataset on a human worker service (for example Mechanical Turk), it is common to give one demonstration of the task. By contrast it is sometimes difficult to communicate the content or format of a task if no examples are given.

**one-shot（1S）**与few-shot类似，只需要一个示例，并搭配语言任务描述。之所以与one-shot、few-shot区分开来，是因为它与人类完成任务的方式最接近。例如，如果要人类生成一个数据集（比如Mechanical Turk），它一般需要给一个示例说明。相反，如果不给出一个示例，人类将较难完成任务。

**Zero-Shot (0S)** is the same as one-shot except that no demonstrations are allowed, and the model is only given a natural language instruction describing the task. This method provides maximum convenience, potential for robustness, and avoidance of spurious correlations (unless they occur very broadly across the large corpus of pre-training data), but is also the most challenging setting. In some cases it may even be difficult for humans to understand the format of the task without prior examples, so this setting is in some cases “unfairly hard”.For example, if someone is asked to “make a table of world records for the 200m dash”, this request can be ambiguous, as it may not be clear exactly what format the table should have or what should be included (and even with careful clarification, understanding precisely what is desired can be difficult). Nevertheless, for at least some settings zero-shot is closest to how humans perform tasks – for example, in the translation example in Figure 2.1, a human would likely know what to do from just the text instruction.

zero-shot(0s)和one-shot类似，它没有示例，模型只需要一个任务描述即可。这个方法提供了最大的便利性、稳健性，并避免出现虚假相关性（除非在预训练中广泛出现该问题），但它也是最有挑战地设置。在一些不提供示例的情况下，即使人类也不好顺畅理解任务，因此这个设置在一些情下是有点困难。例如，如果让某人去完成“make a table of world records for the 200m dash”，这个任务容易产生歧义（即使进行了认真澄清，理解准确也比较困难）。然而，zero-shot是最接近人类完成任务的方式的，例如，在图2.1所示的翻译任务中，人类仅从任务指示中就可以知道如何做。

Figure 2.1 shows the four methods using the example of translating English to French. In this paper we focus on zero-shot, one-shot and few-shot, with the aim of comparing them not as competing alternatives, but as different problem settings which offer a varying trade-off between performance on specific benchmarks and sample efficiency.We especially highlight

the few-shot results as many of them are only slightly behind state-of-the-art fine-tuned models.Ultimately, however, one-shot, or even sometimes zero-shot, seem like the fairest comparisons to human performance,and are important targets for future work.
图2.1展示了四种使用示例进行英译法的方法。在本文，我们重点关注zero-shot、one-shot和few-shot，目的是对比不同的问题设置下的基准测试和样本效率的平衡，而不是方案替代性的。我们尤其重点关注few-shot的结果，因为它们许多都和微调模型最接近。然而，one-shot，有时甚至是few-shot似乎和人类相比更公平，这也是未来的重要研究方向。
Sections 2.1-2.3 below give details on our models, training data, and training process respectively. Section 2.4 discusses the details of how we do few-shot, one-shot, and zero-shot evaluations.
以下2.1到2.3小节给出我们模型的细节，分别是训练数据和训练过程。2.4小节讨论了如何使用few-shot、one-shot和zero-shot进行评估。
2.1 模型和架构
We use the same model and architecture as GPT-2 [RWC+19], including the modified initialization, pre-normalization,and reversible tokenization described therein, with the exception that we use alternating dense and locally banded sparse attention patterns in the layers of the transformer, similar to the Sparse Transformer [CGRS19]. To study the dependence of ML performance on model size, we train 8 different sizes of model, ranging over three orders of magnitude from 125 million parameters to 175 billion parameters, with the last being the model we call GPT-3. Previous work [KMH+20]suggests that with enough training data, scaling of validation loss should be approximately a smooth power law as a function of size; training models of many different sizes allows us to test this hypothesis both for validation loss and for downstream language tasks.
我们使用了和GPT-2一样的模型和架构，包括改变的初始化、预标准化、可逆标记化，但是我们在transforer的层中使用了交替的密集和局部带状稀疏注意力模式，这类似于稀疏Transformer。为研究机器学习表现对模型大小的依赖性，我们训练了8种不同尺寸的模型，从1.25亿到1750亿，跨越了三个数量级，1750亿也就是我们称之为GPT-3的模型。前人的工作表明，只有足够多的数据，验证损失是模型尺寸的平滑幂律函数，训练了这么多不同尺寸的查型也允许我们来验证这个假设，不仅是验证损失上面，还可以下游任务上验证。
Table 2.1 shows the sizes and architectures of our 8 models. Here nparams is the total number of trainable parameters,nlayers is the total number of layers, dmodel is the number of units in each bottleneck layer (we always have the feedforward layer four times the size of the bottleneck layer, dff = 4 ∗ dmodel), and dhead is the dimension of each attention head. All models use a context window of nctx = 2048 tokens. We partition the model across GPUs along both the depth and width dimension in order to minimize data-transfer between nodes. The precise architectural parameters for each model are chosen based on computational efficiency and load-balancing in the layout of models across GPU’s. Previous

work [KMH+20] suggests that validation loss is not strongly sensitive to these parameters within a reasonably broad range.
表2.1显示了我们训练的8个模型的大小和架构。这里nparams是所有训练的参数总量，nlayers是总的层数，dmodel是瓶颈层的单元数，dhead是多头注意力机制中头的数量。所有模型使用的上下文窗口数量为2048。我们根据深度和宽度将模型布在GPU集群上，以此来最小化数据在节点间的传输。每个模型的精确架构参数都是根据在GPU集群上的计算效率和负载均衡来选择的。前人工作表明，验证损失对这些参数不太敏感。

表2.1 我们训练所有模型的大小、架构和学习超参（batch size和学习率）。所有模型的训练tokens都30亿个。

2.2 训练数据

Datasets for language models have rapidly expanded, culminating in the Common Crawl dataset2[RSR+19] constituting nearly a trillion words. This size of dataset is sufficient to train our largest models without ever updating on the same sequence twice. However, we have found that unfiltered or lightly filtered versions of Common Crawl tend to have lower quality than more curated datasets. Therefore, we took 3 steps to improve the average quality of our datasets:(1) we downloaded and filtered a version of CommonCrawl based on similarity to a range of high-quality reference corpora, (2) we performed fuzzy deduplication at the document level, within and across datasets, to prevent redundancy and preserve the integrity of our held-out validation set as an accurate measure of overfitting, and (3) we also added known high-quality reference corpora to the training mix to augment CommonCrawl and increase its diversity.
语言模型的数据集已迅速扩展了，现在的commoncrawl数据集已经有近万亿单词了。这个数据集来训练我们的最大模型其实已经足够了。然而，我们发现未经过滤或仅轻微过滤的common crawl数据相比于精心准备的数据还是有差距。因此，我们采取了三个步骤来提升数据质量：（1）我们下载并过滤了一个CommonCrawl数据，过滤的方法是和一些高质量参考语料进行相似性比对。（2）我们在数据集内和数据集之间进行了文档级的去重。（3）我们将高质量语料加入CommonCrawl数据中，以增加其多样性。
Details of the first two points (processing of Common Crawl) are described in Appendix A. For the third, we added several curated high-quality datasets, including an expanded version of the WebText dataset [RWC+19], collected by scraping links over a longer period of time, and first described in [KMH+20], two internet-based books corpora(Books1 and Books2) and English-language Wikipedia。
前两点（Common Crawl的处理过程）已在附录A中有说明。对于第三点，我们添加了几个精心准备的高质量数据集，包括一个扩展版本的WebText数据集，这是通过抓取一个长时间的链接获得的，它首次出现在[KMH+20]，还有两个互联网书籍语料库（Books1 and Books2），以及一个英语维基百科。

Table 2.2 shows the final mixture of datasets that we used in training. The CommonCrawl data was downloaded from 41 shards of monthly CommonCrawl covering 2016 to 2019, constituting 45TB of compressed plaintext before filtering and 570GB after filtering, roughly equivalent to 400 billion byte-pair-encoded tokens. Note that during training, datasets are not sampled in proportion to their size, but rather datasets we view as higher-quality are sampled more frequently,such that CommonCrawl and Books2 datasets are sampled less than once during training, but the other datasets are sampled 2-3 times. This essentially accepts a small amount of overfitting in exchange for higher quality training data.
表2.2展示了我们最终训练所使用的数据组合。CommonCrawl数据是下载了从2016至2019年41个月度的CommonCrawl数据，最终组成了45T压缩文本，过滤后是570G，这大约相当于4000亿个字节对token。注意，训练时数据从这几份数据集中随机抽取的比例并不与它们的大小成比例，抽取的标准是将高质量的数据比重加大，比如CommonCrawl和Books2r的样本被抽中的次数不超过1，但是其它的有些被抽中了2-3次。这基本上算是接受了一小部分数据的过拟和，来换取高质量的训练数据。

表2.2 训练GPT-3使用的数据

A major methodological concern with language models pretrained on a broad swath of internet data, particularly large models with the capacity to memorize vast amounts of content, is potential contamination of downstream tasks by having their test or development sets inadvertently seen during pre-training. To reduce such contamination, we searched for and attempted to remove any overlaps with the development and test sets of all benchmarks studied in this paper.Unfortunately, a bug in the filtering caused us to ignore some overlaps, and due to the cost of training it was not feasible to retrain the model. In Section 4 we characterize the impact of the remaining overlaps, and in future work we will more aggressively remove data contamination.
大语言模型预训练的一个主要问题是数据污染，也就是在训练时无意看到了测试数据。为了减少数据污染，我们搜索并度图删除了本文所提到的基准测试数据中重叠的部分。不幸的是，过滤时有个bug导致我们漏掉了一些重叠，因成本的原因，再次训练是不可能的。在第四节，我们描述了剩余重叠的影响，在未来，我们也会更加注意消除数据污染的问题。

2.3 训练过程

As found in [KMH+20, MKAT18], larger models can typically use a larger batch size, but require a smaller learning rate. We measure the gradient noise scale during training and use it to guide our choice of batch size [MKAT18]. Table2.1 shows the parameter settings we used. To train the larger models without running out of memory, we use a mixture of model parallelism within each matrix multiply and model parallelism across the layers of the network. All models were trained on V100 GPU’s on part of a high-bandwidth cluster provided by Microsoft. Details of the training process and hyperparameter settings are described in Appendix B.

如[KMH+20, MKAT18]所述，模型越大一般需要更大的batch size，但是需要更小的学习率。我们在训练时评估了梯度噪声，并用它来指导batch size的设定。表2.1列出了我们使用的参数设置。为了训练这个大模型时不至于内存溢出，我们使用了并行策略，这包括矩阵计算方面和网络层级方面。所有的模型都在V100GPU上训练，这是由微软提供的高带宽集群。训练细节和超参在附录B有介绍。

2.4 评估

For few-shot learning, we evaluate each example in the evaluation set by randomly drawing K examples from that task’s training set as conditioning, delimited by 1 or 2 newlines depending on the task. For LAMBADA and Story cloze there is no supervised training set available so we draw conditioning examples from the development set and evaluate on the test set. For Winograd (the original, not SuperGLUE version) there is only one dataset, so we draw conditioning examples directly from it.

对于few-shot学习，我们从训练集中随机抽取k个样例，然后在评估集上评估每个样例，根据任务的不同，k个样例之间有1或2个分隔符。对于LAMBADA和Story cloze，因为没有训练集，因此从开发集中抽取样例作为条件，并在测试集上评估。对于Winograd（原始的，不是SuperGlue版本），仅有一个数据集，因此我们直接从中抽取样例来作为条件。

K can be any value from 0 to the maximum amount allowed by the model’s context window, which is nctx = 2048 for all models and typically fits 10 to 100 examples. Larger values of K are usually but not always better, so when a separate development and test set are available, we experiment with a few values of K on the development set and then run the best value on the test set. For some tasks (see Appendix G) we also use a natural language prompt in addition to(or for K = 0, instead of) demonstrations.

k可以从0到最大值，最大值由上下文窗口2048决定，通常对应10至100个样例。一般来说，k越大效果越好，但也不总是这样，所以当有独立的开发集和测试集时，我们在开发集上实验不同k值样例，在测试集上评估，得到佳的k值。对一些任务（见附录G），我们也会使用prompt，用来补充、或替代样例。

On tasks that involve choosing one correct completion from several options (multiple choice), we provide K examplesof context plus correct completion, followed by one example of context only, and compare the LM likelihood of each completion. For most tasks we compare the per-token likelihood (to normalize for length), however on a small number of datasets (ARC, OpenBookQA, and RACE) we gain additional benefit as measured on the development set by normalizing by the unconditional probability of each completion, by computing P (completion|context)P (completion|answer context) , where answer context is the string "Answer: " or "A: " and is used to prompt that the completion should be an answerbut is otherwise generic.（此段落较难翻译，暂时搁置）

On tasks that involve binary classification, we give the options more semantically meaningful names (e.g. “True” or “False” rather than 0 or 1) and then treat the task like multiple

choice; we also sometimes frame the task similar to whatis done by [RSR+19] (see Appendix G) for details.
在一些二分类任务上，我们会给选项更具意义的名称（比如“True”或“False”，而不是0或1），然后将其当作多选任务。我们有时也将任务的框架设为与[RSR+19]一样。
On tasks with free-form completion, we use beam search with the same parameters as [RSR+19]: a beam width of 4 and a length penalty of α = 0.6. We score the model using F1 similarity score, BLEU, or exact match, depending on what is standard for the dataset at hand。
在自由文本输出这样的任务中，我们使用与[RSR+19]相同的参数进行束搜索：束宽为4，长度惩罚为α = 0.6。我们使用F1相似度评分、BLEU或精准匹配来评估，具体使用哪一种取决于手头数据集的标准。
Final results are reported on the test set when publicly available, for each model size and learning setting (zero-, one-,and few-shot). When the test set is private, our model is often too large to fit on the test server, so we report results onthe development set. **We do submit to the test server on a small number of datasets (SuperGLUE, TriviaQA, PiQa) where we were able to make submission work, and we submit only the 200B few-shot results, and report developmentset results for everything else.（加粗这一块暂时搁置，不理解）**
如果测试集是公开的，我们都对测试集做了报告，这包括了模型大小维度和学习设置（zero-、one-、few-shot）。当测试集是私有的时，我们的模型因为太大无法在其服务器上部署使用，所以我们报告了它的验证集结果。
3 结果
In Figure 3.1 we display training curves for the 8 models described in Section 2. For this graph we also include 6 additional extra-small models with as few as 100,000 parameters. As observed in [KMH+20], language modeling performance follows a power-law when making efficient use of training compute. After extending this trend by two more orders of magnitude, we observe only a slight (if any) departure from the power-law. One might worry that these improvements in cross-entropy loss come only from modeling spurious details of our training corpus. However, we will see in the following sections that improvements in cross-entropy loss lead to consistent performance gains across abroad spectrum of natural language tasks.
在图3.1中，我们展示了第2节中所说的八个模型的训练曲线。在这个图中，我们还增加了6个小模型，最小的有10万参数。正如[KMH+20]中所提到的，语言模型的表现跟算力呈幂律关系。算力扩大2个数量级，幂律仅有微小偏移。有人可能会担心，这些交叉熵损失的改进是因为存在训练过程的虚假细节。然而，接下来的章节中将会发现，交叉熵损失的改善在广泛的nlp任务中有一致的效果提升。

图3.1 验证损失随着算力增加稳定提升

Below, we evaluate the 8 models described in Section 2 (the 175 billion parameter parameter GPT-3 and 7 smaller models) on a wide range of datasets. We group the datasets into 9 categories representing roughly similar tasks.
下面，我们在一系列数据集上评估了第二节提到的8个模型（1750亿的GPT-3以及7个小模型）。我们将这些数据集分成9类。
In Section 3.1 we evaluate on traditional language modeling tasks and tasks that are similar to language modeling,such as Cloze tasks and sentence/paragraph completion tasks. In Section 3.2 we evaluate on "closed book" question answering tasks: tasks which require using the information stored in the model's parameters to answer general knowledge questions. In Section 3.3 we evaluate the model's ability to translate between languages (especially one-shot and few-shot). In Section 3.4 we evaluate the model's performance on Winograd Schema-like tasks. In Section 3.5 we evaluate on datasets that involve commonsense reasoning or question answering. In Section 3.6 we evaluate on reading comprehension tasks, in Section 3.7 we evaluate on the SuperGLUE benchmark suite, and in 3.8 we briefly explore NLI. Finally, in Section 3.9, we invent some additional tasks designed especially to probe in-context learning abilities –these tasks focus on on-the-fly reasoning, adaptation skills, or open-ended text synthesis. We evaluate all tasks in the few-shot, one-shot, and zero-shot settings.
在3.1节，我们评估了传统的语言模型任务以及类似于语言模型的任务，比如完形填空和句子/段落补齐任务。在3.2节，我们评估了闭书问答任务：也就是需要模型参数中存储的信息来回答通用问题。在3.3节，评估了翻译任务（尤其是在one-shot和few-shot）的情况下。在3.4节，我们评估了Winograd Schema-like任务（考验理解能力的任务）。在3.5节，我们评估了常识推理任务。在3.6节，我们评估了阅读理解任务。在3.7节，我们评估了SuperGLUE基线任务，在3.8节，我们简要探索了自然语言推理任务。最后，在3.9节，我们专门设计了一些附加任务来探索上下文学习能力，这些能力聚焦在即时推理，适应性能力或开放式文本合成方面。我们评估这些任务都用了few-shot、one-shot及zero-shot这些设置。
3.1 语言模型、完形填空和补全任务
In this section we test GPT-3's performance on the traditional task of language modeling, as well as related tasks that involve predicting a single word of interest, completing a sentence or paragraph, or choosing between possible completions of a piece of text.
在本小节，我们测试了GPT-3在传统语言模型任务上的表现，还有相关的任务，比如预测感兴趣的词，补全句子或段落，或者从可能的补全段落中选择一个。
3.1.1 语言模型
We calculate zero-shot perplexity on the Penn Tree Bank (PTB) [MKM+94] dataset measured in [RWC+19]. We omit the 4 Wikipedia-related tasks in that work because they are entirely contained in our training data, and we also omit the one-billion word benchmark due to a high fraction of the dataset being contained in our training set. PTB escapes these issues due

to predating the modern internet. Our largest model sets a new SOTA on PTB by a substantial margin of 15 points, achieving a perplexity of 20.50. Note that since PTB is a traditional language modeling dataset it does not have a clear separation of examples to define one-shot or few-shot evaluation around, so we measure only zero-shot.
我们在PTB数据上评估了zero-shot困惑度。有些语言模型数据因其全部或大部分包含在我们的训练数据中，因并没有采用，比如4个维基百科相关的任务，还有one-billion word基准。PTB之所以可以避免这些问题，是因为它早于现代互联网之前就有了。我们最大的模型在PTB超越了现有SOTA15个百分点，困惑度为20.5。因为PTB是较传统的数据，因此并没有有效的分隔符来定义one-shot和few-shot评估，故我们只用了zero-shot来评估。

3.1.2 LAMBADA

The LAMBADA dataset [PKL+16] tests the modeling of long-range dependencies in text – the model is asked to predict the last word of sentences which require reading a paragraph of context. It has recently been suggested that the continued scaling of language models is yielding diminishing returns on this difficult benchmark. [BHT+20] reflect on the small 1.5% improvement achieved by a doubling of model size between two recent state of the art results ([SPP+19] and [Tur20]) and argue that "continuing to expand hardware and data sizes by orders of magnitude is not the pathforward". We find that path is still promising and in a zero-shot setting GPT-3 achieves 76% on LAMBADA, a gain of 8% over the previous state of the art.
LAMBADA数据集可以用来测试长文本依赖，模型需要阅读文本段落，然后才能预测最后一个词。近期有研究称，扩大语言模型规模对这一基准测试的效果提升没那么明显。[BHT+20]回顾了近期的两个先进的结果[SPP+19] and [Tur20]，发现模型大小翻倍，但效果仅提升了1.5%，因此得到了结论"继续扩大硬件和数据大小的量级不是一个好的方向"。我们发现这条路依然充满希望，在zero-shot上，GPT-3在LAMBADA上获得了76%的结果，较之前最先进的结果提升了8%。
LAMBADA is also a demonstration of the flexibility of few-shot learning as it provides a way to address a problem that classically occurs with this dataset. Although the completion in LAMBADA is always the last word in a sentence, a standard language model has no way of knowing this detail. It thus assigns probability not only to the correct ending but also to other valid continuations of the paragraph. This problem has been partially addressed in the past with stop-wordfilters [RWC+19] (which ban "continuation" words). The few-shot setting instead allows us to "frame" the task as a cloze-test and allows the language model to infer from examples that a completion of exactly one word is desired. We use the following fill-in-the-blank format:

Alice was friends with Bob. Alice went to visit her friend ———. → Bob

George bought some baseball equipment, a ball, a glove, and a ——. →

LAMBADA也证明了few-shot学习的灵活性，它提供了解决这个数据集问题的一个方法。尽管LAMBADA要补充的是一个句子的最后一个词，标准语言模型是不知道这个细节的。标准语言模型不

仅会给出最后一个词的概率，也会给出往后延续的其它词的概率。这个问题通过设置停止命令可以部分解决。few-shot学习的解决方法是将任务定义为完形填空，让语言模型参考示例预测最后一个词。下面是一个例子：

Alice was friends with Bob. Alice went to visit her friend ———. → Bob

George bought some baseball equipment, a ball, a glove, and a ——. →

When presented with examples formatted this way, GPT-3 achieves 86.4% accuracy in the few-shot setting, an increase of over 18% from the previous state-of-the-art. We observe that few-shot performance improves strongly with model size. While this setting decreases the performance of the smallest model by almost 20%, for GPT-3 it improves accuracy by 10%. Finally, the fill-in-blank method is not effective one-shot, where it always performs worse than the zero-shot setting. Perhaps this is because all models still require several examples to recognize the pattern.

当以这种方式展示示例时，GPT-3在few-shot上提升了86.4%，比之前最好水平提升了18%。我们发现模型越大，few-shot效果越好。然而在最小模型上，few-shot效果却下降了20%，GPT-3上是提升了10%。最后，填空这种方法对于one-shot不太适用，它比zero-shot还差。也许模型仍然是需要多个示例来识别模式吧。

One note of caution is that an analysis of test set contamination identified that a significant minority of the LAMBADAdataset appears to be present in our training data – however analysis performed in Section 4 suggests negligible impacton performance.

图3.2 在LAMBADA，few-shot的准确率提升很多。

One note of caution is that an analysis of test set contamination identified that a significant minority of the LAMBADA dataset appears to be present in our training data – however analysis performed in Section 4 suggests negligible impact on performance.

需要注意地一点是，对测试集的数据污染分析发现，LAMBADA数据集有相当一部分出现在训练数据中，然而，我们在第节的分析结果表，对效果影响不大。

3.1.3 HellaSwag

The HellaSwag dataset [ZHB+19] involves picking the best ending to a story or set of instructions. The examples were adversarially mined to be difficult for language models while remaining easy for humans (who achieve 95.6% accuracy).GPT-3 achieves 78.1% accuracy in the one-shot setting and 79.3% accuracy in the few-shot setting, outperforming the 75.4% accuracy of a fine-tuned 1.5B parameter language model [ZHR+19] but still a fair amount lower than the overall SOTA of 85.6% achieved by the fine-tuned multi-task model ALUM.

HellaSwag数据集是从一个故事或几个指令中选一个最佳结局。这些数据都经过了对抗挖掘，对语言模型来说有些难度，但是人类来说并不复杂（人类可以达到95.6%准确率）。GPT-3在one-shot上达

到了78.1%，在few-shot上达到了79.3%，比1.5B的微调模型[ZHR+19]的75.4%是高一些，但是距多任务模型ALUM的85.6%还差很远。

3.1.4 StoryCloze

We next evaluate GPT-3 on the StoryCloze 2016 dataset [MCH+16], which involves selecting the correct ending sentence for five-sentence long stories. Here GPT-3 achieves 83.2% in the zero-shot setting and 87.7% in the few-shot setting (with K = 70). This is still 4.1% lower than the fine-tuned SOTA using a BERT based model [LDL19] but improves over previous zero-shot results by roughly 10%.

我们接下来评估了StoryCloze 2016数据集，该数据集是从大约五个句子的故事中找到正确的结束句子。GPT-3在one-shot下达到了83.2%，在few-shot下达到了87.7%（K=70）。这跟之前基于BERT的SPTA结论相比还是低了4.1%，但是比我们之前的zero-shot高了有10%左右。

3.2 闭书问答

In this section we measure GPT-3’s ability to answer questions about broad factual knowledge. Due to the immense amount of possible queries, this task has normally been approached by using an information retrieval system to find relevant text in combination with a model which learns to generate an answer given the question and the retrieved text. Since this setting allows a system to search for and condition on text which potentially contains the answer itis denoted “open-book”. [RRS20] recently demonstrated that a large language model can perform surprisingly well directly answering the questions without conditioning on auxilliary information. They denote this more restrictive evaluation setting as “closed-book”. Their work suggests that even higher-capacity models could perform even better and we test this hypothesis with GPT-3. We evaluate GPT-3 on the 3 datasets in [RRS20]: Natural Questions [KPR+19],WebQuestions [BCFL13], and TriviaQA [JCWZ17], using the same splits. Note that in addition to all results being in the closed-book setting, our use of few-shot, one-shot, and zero-shot evaluations represent an even stricter setting than previous closed-book QA work: in addition to external content not being allowed, fine-tuning on the Q&A dataset itself is also not permitted.

在本小节，我们评估GPT-3在广泛的事实问题上的问答能力。因为查询数据巨大，这个任务的做法一般是先检索相关信息，然后根据检索到的内容和问题使用大模型答案。因为这个设置允许系统搜索文本，并基于文本回答，这个文本潜在地包含了答案，因此被称作“开卷”。[RRS20]近期证明了，更高容量的模型可以直接回答问题，而不用基于辅助信息。他们将此更加严格的设置称为“闭卷”。我们在三个数据集上评估了GPT-3，Natural Questions [KPR+19],WebQuestions [BCFL13], and TriviaQA [JCWZ17]，其中数据集划分方法一样。请注意，除了所有结果都在闭卷设置下，我们使用的少样本、一次样本和零样本评估代表了比以前的闭卷 QA 工作更严格的设置：除了不允许使用外部内容外，还不允许在问答数据集本身上进行微调。

The results for GPT-3 are shown in Table 3.3. On TriviaQA, we achieve 64.3% in the zero-shot setting, 68.0% in the one-shot setting, and 71.2% in the few-shot setting. The zero-shot result

already outperforms the fine-tuned T5-11B by 14.2%, and also outperforms a version with Q&A tailored span prediction during pre-training by 3.8%. The one-shot result improves by 3.7% and matches the SOTA for an open-domain QA system which not only fine-tunes but also makes use of a learned retrieval mechanism over a 15.3B parameter dense vector index of 21M documents [LPP+20].GPT-3’s few-shot result further improves performance another 3.2% beyond this.
GPT-3的评估结果见表3.3。在TriviaQA上，zero-shot上得到了64.3%，在one-shot上得到了68%，在few-shot上得到了71.2%。zero-shot的结果已经超越了微调的T5-11B模型14.2%个点，而且超越了预训练期间针对问进行了跨度预测的T5版本3.8%。one-shot结果相较于zero-shot提升了3.7%，且和RAG持平，RAG不仅有微调，而且使用了21M文档生成的15.3B参数的密集向量索引。GPT-3的few-shot结果进一步比one-shot提升3.2%。

表 3.3: 三个开源QA任务的结果

On WebQuestions (WebQs), GPT-3 achieves 14.4% in the zero-shot setting, 25.3% in the one-shot setting, and 41.5%in the few-shot setting. This compares to 37.4% for fine-tuned T5-11B, and 44.7% for fine-tuned T5-11B+SSM,which uses a Q&A-specific pre-training procedure. GPT-3 in the few-shot setting approaches the performance of state-of-the-art fine-tuned models. Notably, compared to TriviaQA, WebQS shows a much larger gain from zero-shot to few-shot (and indeed its zero-shot and one-shot performance are poor), perhaps suggesting that the WebQs questions and/or the style of their answers are out-of-distribution for GPT-3. Nevertheless, GPT-3 appears able to adapt to this distribution, recovering strong performance in the few-shot setting.
在WebQuestions，GPT-3在zero-shot下达到了14.4%，在one-shot下达到了25.3%，在few-shot下达到了41.5%。作为对比，T5-11B为37.4%，T5-11B+SSM为44.7%。GPT-3在few-shot情况下达到了微调的最好模型的水平。值得注意的是，与TriviaQA相比，WebQS的zero-shot到few-shot之间差异很大（zero-shot和one-shot确实效果很差），也许WebQs的问题和答案是GPT-3分布外的。尽管如此，GPT-3还是可以通过few-shot来适应这个分布，恢复比较好的效果。
On Natural Questions (NQs) GPT-3 achieves 14.6% in the zero-shot setting, 23.0% in the one-shot setting, and 29.9% in the few-shot setting, compared to 36.6% for fine-tuned T5 11B+SSM. Similar to WebQS, the large gain from zero-shot to few-shot may suggest a distribution shift, and may also explain the less competitive performance compared toTriviaQA and WebQS. In particular, the questions in NQs tend towards very fine-grained knowledge on Wikipedia specifically which could be testing the limits of GPT-3’s capacity and broad pretraining distribution.
在Natural Questions上，GPT-3在zero-shot达到了14.6%，在one-shot上达到了23%，在few-shot上达到了29.9%，与之对应，T5 11B+SSM为36.6%。与WebQS类似，zero-shot到few-shot的较大

差异可能揭示了分布的变化。尤其是，NQs的问题倾向于维基百科的细粒度知识，这可能会考验GPT-3的能力和预训练分布的极限。

Overall, on one of the three datasets GPT-3’s one-shot matches the open-domain fine-tuning SOTA. On the other two datasets it approaches the performance of the closed-book SOTA despite not using fine-tuning. On all 3 datasets, we find that performance scales very smoothly with model size (Figure 3.3 and Appendix H Figure H.7), possibly reflecting the idea that model capacity translates directly to more ‘knowledge’ absorbed in the parameters of the model.

总之，在其中一个数据集上，GPT-3的one-shot和开卷微调SOTA持平。在另个两个数据集上，尽管没用微调，也达到了闭卷SOTA的效果。在这三个数据集上，我们发现效果与模型大小是成正比的（图3.3和附录Hl图H.7），印证了，模型参数就是知识。

图3.3 在TriviaQAGPT-3的效果随着模型大小而稳定提升，表明语言模型随着大小的提升，持续地吸引知识。

3.3 翻译

For GPT-2 a filter was used on a multilingual collection of documents to produce an English only dataset due to capacity concerns. Even with this filtering GPT-2 showed some evidence of multilingual capability and performed non-trivially when translating between French and English despite only training on 10 megabytes of remaining French text. Since we increase the capacity by over two orders of magnitude from GPT-2 to GPT-3, we also expand the scope of the training dataset to include more representation of other languages, though this remains an area for further improvement. As discussed in 2.2 the majority of our data is derived from raw Common Crawl with only quality-based filtering. Although GPT-3’s training data is still primarily English (93% by word count), it also includes 7% of text in other languages.These languages are documented in the supplemental material. In order to better understand translation capability, we also expand our analysis to include two additional commonly studied languages, German and Romanian.

对于GPT-2，因为容量问题，数据集是使用了一个英语过滤器。尽管经过了过滤，GPT-2在法对英翻译上表现非凡，其中训练数据中仅有10M字节法语文本。因GPT-2到GPT-3的模型容量扩大了两个数量级，我们也将训练数据囊括了更多的其它语种，尽管这还是一个需要提升的领域。如2.2小节讨论的，大部分数据都是来自Common Crawl并且仅经过一次质量层面的过滤。尽管GPT-3的训练数据仍然主要是英语，它还有7%的其它语种。这些语种在补充材料中有记录。为了更好地理解翻译能力，我们还扩大了分析范围，即常用的语种，德语和罗马尼亚语。

Existing unsupervised machine translation approaches often combine pretraining on a pair of monolingual datasets with back-translation [SHB15] to bridge the two languages in a controlled way. By contrast, GPT-3 learns from a blend of training data that mixes many languages to gether in a natural way, combining them on a word, sentence,and document

level. GPT-3 also uses a single training objective which is not customized or designed for any task in particular. However, our one / few-shot settings aren’ t strictly comparable to prior unsupervised work since they make use of a small amount of paired examples (1 or 64). This corresponds to up to a page or two of in-context training data.
现有的无监督机器翻译方法通常将预训练应用于一对单语数据集，然后通过反向翻译以受控方式连接两种语言。然而，GPT-3的方式不同，它是从多语种训练数据中学习，融合了词、句、段落几种级别。GPT-3还使用了一个统一的目标函数，并没有什对特定任务来设计。然而，我们的one-shot、few-shot设置并不能严格的和之前的无监督工作相比，我们是使用了少量的成对示例（1 or 64），这相当于训练数据中的一页或两页的上下文。
Results are shown in Table 3.4. Zero-shot GPT-3, which only receives on a natural language description of the task,still underperforms recent unsupervised NMT results. However, providing only a single example demonstration for each translation task improves performance by over 7 BLEU and nears competitive performance with prior work.GPT-3 in the full few-shot setting further improves another 4 BLEU resulting in similar average performance to prior unsupervised NMT work. GPT-3 has a noticeable skew in its performance depending on language direction. For the three input languages studied, GPT-3 significantly outperforms prior unsupervised NMT work when translating into English but underperforms when translating in the other direction. Performance on En-Ro is a noticeable outlier at over 10 BLEU worse than prior unsupervised NMT work. This could be a weakness due to reusing the byte-level BPE tokenizer of GPT-2 which was developed for an almost entirely English training dataset. For both Fr-En and De-En,few shot GPT-3 outperforms the best supervised result we could find but due to our unfamiliarity with the literature and the appearance that these are un-competitive benchmarks we do not suspect those results represent true state of the art.For Ro-En, few shot GPT-3 performs within 0.5 BLEU of the overall SOTA which is achieved by a combination of unsupervised pretraining, supervised finetuning on 608K labeled examples, and backtranslation [LHCG19b].
结果见表3.4。zero-shot GPT-3仅通过任务说明，效果不如近期的NMT结果。然而，增加一个示例后，效果提升了7个BLEU，和之前的工作相比也有一定竞争力。使用了充分的few-shot后，近一步提升了4个BLEU，平均水平和之前NMT工作效果差不多。GPT-3在不同的翻译方向上有较大的偏差。这三个语言作为输入，GPT-3的效果明显超越了先前的无监督NMT，当反过来时，则不如先前的NMT。En-Ro的表现明显低于先前NMT10个BLEU。由于重复使用为几乎全英语训练数据集的字节级tokenizer，这可能是一个弱点。对于法语-英语和德语-英语，few-shot GPT-3优于我们能找到的最佳监督结果，但由于我们对文献的不熟悉以及这些结果看起来不具有竞争力，我们不怀疑这些结果代表了真正的最新技术水平。对于罗马尼亚语-英语，few-shot GPT-3的表现与之前SOTA相差0.5个点，这是由无监督预训练，608K标记示例的监督微调和反向翻译[LHCG19b]所实现。

表3.4 在其它语言翻译成英语方面，few-shot GPT-3超越了无监督神经网络模型5个BLEU

Finally, across all language pairs and across all three settings (zero-, one-, and few-shot), there is a smooth trend of improvement with model capacity. This is shown in Figure 3.4 in the case of few-shot results, and scaling for all three settings is shown in Appendix H.
最后，在所有的语言对和所有的设置（zero-shot、one-shot、few-shot）中，都有一个随模型大小而稳定提升模型表现的规律。详情见图3.4，这是few-shot的结果，更详细的结果见附录H。

图3.4 在模型大小改变的情况下，6个语言对的翻译表现

3.4 Winograd-Style任务

The Winograd Schemas Challenge [LDM12] is a classical task in NLP that involves determining which word a pronoun refers to, when the pronoun is grammatically ambiguous but semantically unambiguous to a human. Recently fine-tuned language models have achieved near-human performance on the original Winograd dataset, but more difficult versions such as the adversarially-mined Winogrande dataset [SBBC19] still significantly lag human performance. We test GPT-3’s performance on both Winograd and Winogrande, as usual in the zero-, one-, and few-shot setting.
Winograd模式挑战[LDM12]是一个经典的任务，主要是判断代词指的是哪一个词，同时代词在语法上是有歧义的，但是对人类来说并不难。近期的微调模型在Winograd dataset上已接近了人类水平，但是在更难的版本上，比如对抗性挖掘的Winogrande dataset[SBBC19]依然落后于人类。我们在Winograd and Winogrande都测试了GPT-3的表现，和以前一样，使zero-shot、one-shot和few-shot设置。

On Winograd we test GPT-3 on the original set of 273 Winograd schemas, using the same “partial evaluation” method described in [RWC+19]. Note that this setting differs slightly from the WSC task in the SuperGLUE benchmark, which is presented as binary classification and requires entity extraction to convert to the form described in this section. On Winograd GPT-3 achieves 88.3%, 89.7%, and 88.6% in the zero-shot, one-shot, and few-shot settings, showing no clear in-context learning but in all cases achieving strong results just a few points below state-of-the-art and estimated human performance. We note that contamination analysis found some Winograd schemas in the training data but this appears to have only a small effect on results (see Section 4).
在Winograd上，我们测试了原始的273个Winograd主题，使用的是“部分评估”方法。注意，这个设置与SuperGLUE基准测试中的WSC略有不同，WSC是当作二分类任务，且需要实体抽取以转成本节所描述的形式。在Winograd，GPT-3在zero-shot、one-shot和few-shot分别达到了88.3%、89.7%和88.6%，结果表明，上下文学习并不明显，但结果很好，仅比SOTA和人类表现差几点。我们还做了数据污染分析，发现在训练数据中存在一些Winograd主题，但是这对结果影响很小（见第四节）

On the more difficult Winogrande dataset, we do find gains to in-context learning: GPT-3 achieves 70.2% in the zero-shot setting, 73.2% in the one-shot setting, and 77.7% in the few-shot setting. For comparison a fine-tuned RoBERTA model achieves 79%, state-of-the-art is

84.6% achieved with a fine-tuned high capacity model (T5), and human performance on the task as reported by [SBBC19] is 94.0%.
在更复杂的Winogrande数据上，我们发现上下文学习有效果提升：GTP-3在zero-shot上达到了70.2%，在one-shot上达到了73.2%，在few-shot上达到了77.7%。作为对比，微调模型RoBERTA达到了79%，SOTA是一个微调模型T5，为84.6%，人类水平是94%。
3.5 常识推理
Next we consider three datasets which attempt to capture physical or scientific reasoning, as distinct from sentence completion, reading comprehension, or broad knowledge question answering. The first, PhysicalQA (PIQA) [BZB+19],asks common sense questions about how the physical world works and is intended as a probe of grounded understanding of the world. GPT-3 achieves 81.0% accuracy zero-shot, 80.5% accuracy one-shot, and 82.8% accuracy few-shot(the last measured on PIQA’ s test server). This compares favorably to the 79.4% accuracy prior state-of-the-art of a fine-tuned RoBERTa. PIQA shows relatively shallow scaling with model size and is still over 10% worse than human performance, but GPT-3’ s few-shot and even zero-shot result outperform the current state-of-the-art. Our analysis flagged PIQA for a potential data contamination issue (despite hidden test labels), and we therefore conservatively mark the result with an asterisk. See Section 4 for details.
接下来，我们来考虑三个关于物理或科学方面的常识推理数据集。第一个是 PhysicalQA (PIQA)，它会问一些关于这个物理世界如何运作的问题，旨在探讨对这个世界的理解。GPT-3在zero-shot下得到了81%，在one-shot下得到了80.5%，在few-shot下得到了82.8%（最后一个是在PIQA的测试服务器上测的）。而之前最好的微调模型RoBERTa是79.4%。随着模型规模变大，效果并没有显著提升，且离人类表现有超过10%的差距，但是GPT-3的few-shot超越了SOTA，即使zero-shot也超越了。我们分析到PIQA有数据污染现象（尽管隐藏了测试标签），所以我们谨慎地将结果标了星。详见第四节。

表3.6 GPT-3在三个常识推理任务（PIQA, ARC, and OpenBookQA）上的结果

Figure 3.6: GPT-3在zero-shot、one-shot和few-shot设置下，在PIQA上的表现
ARC [CCE+18] is a dataset of multiple-choice questions collected from 3rd to 9th grade science exams. On the “Challenge” version of the dataset which has been filtered to questions which simple statistical or information retrieval methods are unable to correctly answer, GPT-3 achieves 51.4% accuracy in the zero-shot setting, 53.2% in the one-shot setting, and 51.5% in the few-shot setting. This is approaching the performance of a fine-tuned RoBERTa baseline(55.9%) from UnifiedQA [KKS+20]. On the “Easy” version of the dataset (questions which either of the mentioned baseline approaches answered correctly), GPT-3 achieves 68.8%, 71.2%, and 70.1% which slightly exceeds a fine-tuned RoBERTa baseline from [KKS+20]. However, both of these results are still much worse than the overall

SOTAs achieved by the UnifiedQA which exceeds GPT-3’s few-shot results by 27% on the challenge set and 22% on the easy set.
ARC是3到9年级的科学考试的多选题数据集。有挑战版和简单版，挑战版是通过简单地统计和检索无法得到答案。在挑战版，GPT-3在zero-shot上得到了51.4%，在one-shot上得到了53.2%，在few-shot上得到了51.5%。这接近于UnifiedQA 中的RoBERTa基线(55.9%)。在简单版中，GPT-3在zero-shot上68.8%，在one-shot上71.2%，在few-shot上70.1%，略高于RoBERTa基线。然而，不管挑战版还是简单版，这个结果比UnifiedQA中的整体SOTA都差很多，挑战版比few-shot高27%，简单版比few-shot高22%。
On OpenBookQA [MCKS18], GPT-3 improves significantly from zero to few shot settings but is still over 20 points short of the overall SOTA. GPT-3’s few-shot performance is similar to a fine-tuned BERT Large baseline on the leaderboard.
在OpenBookQA [MCKS18]，GPT-3从zero-shot到few-shot是有提升，但是和总体SOTA相比有20个点的差距。GPT-3的few-shot性能和排行榜上的bert large基类似。
Overall, in-context learning with GPT-3 shows mixed results on commonsense reasoning tasks, with only small and inconsistent gains observed in the one and few-shot learning settings for both PIQA and ARC, but a significant improvement is observed on OpenBookQA. GPT-3 sets SOTA on the new PIQA dataset in all evaluation settings.
总之，GPT-3在常识推理方面的上下文学习中表现不一，在 PIQA and ARC上，在one-shot到few-shot上有小的提升，但是在OpenBookQA上有大的提升。GPT-3在PIQA数据集上所有的评估设置中都刷新了SOTA。
3.6 阅读理解
Next we evaluate GPT-3 on the task of reading comprehension. We use a suite of 5 datasets including abstractive,multiple choice, and span based answer formats in both dialog and single question settings. We observe a wide spread in GPT-3’s performance across these datasets suggestive of varying capability with different answer formats. In general we observe GPT-3 is on par with initial baselines and early results trained using contextual representations on each respective dataset.
接下来，我们评估GPT-3在阅读理解任务上的表现。我们使用了5个数据集，包括抽象回答、多项选择和基于跨度的回答格式，这些数据集涵盖了对话和单个问题的设置。我们观察到GPT-3在这些数据集上的表现差异很大，这表明它在不同的回答格式上具有不同的能力。总体而言，我们观察到GPT-3与基线以及早期用上下文表示的结果相当。
GPT-3 performs best (within 3 points of the human baseline) on CoQA [RCM19] a free-form conversational dataset and performs worst (13 F1 below an ELMo baseline) on QuAC [CHI+18] a dataset which requires modeling structured dialog acts and answer span selections of teacher-student interactions. On DROP [DWD+19], a dataset testing discrete reasoning and numeracy in the context of reading comprehension, GPT-3 in a few-shot setting outperforms the fine-tuned BERT baseline from the original paper but is still well

below both human performance and state-of-the-art approaches which augment neural networks with symbolic systems [RLL+19]. On SQuAD 2.0 [RJL18], GPT-3 demonstrates its few-shot learning capabilities, improving by almost 10 F1 (to 69.8) compared to a zero-shot setting. This allows it to slightly outperform the best fine-tuned result in the original paper. On RACE [LXL+17], a multiple choice dataset ofmiddle school and high school english examinations, GPT-3 performs relatively weakly and is only competitive with the earliest work utilizing contextual representations and is still 45% behind SOTA.
GPT-3是一种语言模型，它在不同的数据集上表现不同。在CoQA [RCM19]这个自由形式的对话数据集上，GPT-3的表现最好，仅比人类基线低3分。在QuAC [CHI+18]这个需要构建结构化对话行为和教师-学生交互的数据集上，GPT-3的表现最差，比ELMo基线低13 F1。在DROP [DWD+19]这个测试离散推理和阅读理解背景下的数字能力的数据集上，GPT-3在少样本情况下优于原始论文中的经过微调的BERT基线，但仍远低于人类表现和采用符号系统增强神经网络的最先进方法。在SQuAD 2.0 [RJL18]上，GPT-3展示了其少样本学习能力，相比零样本情况下提高了近10 F1（达到69.8）。这使它略微优于原始论文中的最佳微调结果。在RACE [LXL+17]上，这是一个中学和高中英语考试的多项选择数据集，GPT-3表现相对较弱，只能与使用上下文表示的最早工作效果相近，并且仍然落后于最先进的技术水平45%。
3.7 SuperGLUE
In order to better aggregate results on NLP tasks and compare to popular models such as BERT and RoBERTa in a more systematic way, we also evaluate GPT-3 on a standardized collection of datasets, the SuperGLUE benchmark[WPN+19] [WPN+19] [CLC+19] [DMST19] [RBG11] [KCR+18] [ZLL+18] [DGM06] [BHDD+06] [GMDD07][BDD+09] [PCC18] [PHR+18]. GPT-3’s test-set performance on the SuperGLUE dataset is shown in Table 3.8. In the few-shot setting, we used 32 examples for all tasks, sampled randomly from the training set. For all tasks except WSC and MultiRC, we sampled a new set of examples to use in the context for each problem. For WSC and MultiRC, we used the same set of randomly drawn examples from the training set as context for all of the problems we evaluated.
为了更好地聚合NLP任务的结果并以更系统的方式与流行模型（如BERT和RoBERTa）进行比较，我们还在标准化的数据集SuperGLUE基准测试[WPN+19] [WPN+19] [CLC+19] [DMST19] [RBG11] [KCR+18] [ZLL+18] [DGM06] [BHDD+06] [GMDD07][BDD+09] [PCC18] [PHR+18]上评估了GPT-3。GPT-3在SuperGLUE数据集上的测试集表现如表3.8所示。在few-shot情况下，我们对所有任务使用了32个样例，这些样例是从训练集中随机抽取的。除WSC和MultiRC外，我们都会抽取一组新的示例来用于每个问题的上下文。对于WSC和MultiRC，我们使用了从训练集中随机抽取的相同示例集作为我们评估的所有问题的上下文。

表3.8 在SuperGLUE上，GPT-3和微调模型及SOTA对比

We observe a wide range in GPT-3’s performance across tasks. On COPA and ReCoRD GPT-3 achieves near-SOTA performance in the one-shot and few-shot settings, with COPA falling

only a couple points short and achieving second place on the leaderboard, where first place is held by a fine-tuned 11 billion parameter model (T5). On WSC,performance is still relatively strong, achieving 80.1% in the few-shot setting (note that GPT-3 achieves 88.6% on theoriginal Winograd dataset as described in Section 3.4). On BoolQ, MultiRC, and RTE, performance is reasonable,roughly matching that of a fine-tuned BERT-Large. On CB, we see signs of life at 75.6% in the few-shot setting.

我们观察到GPT-3在不同任务上的表现差异很大。在COPA和ReCoRD上，GPT-3在一次和少次样本情况下均取得了接近SOTA的表现，其中COPA仅落后几分，并在排行榜上获得了第二名，而第一名由经过微调的110亿参数模型（T5）获得。在WSC上，性能仍然相对较强，在few-shot情况下达到了80.1%（请注意，GPT-3在原始Winograd数据集上的表现为88.6%，如第3.4节所述）。在BoolQ、MultiRC和RTE上，性能合理，大致与经过微调的BERT-Large相匹配。在CB上，我们在few-shot情况下达到了75.6%。

WiC is a notable weak spot with few-shot performance at 49.4% (at random chance). We tried a number of different phrasings and formulations for WiC (which involves determining if a word is being used with the same meaning in two sentences), none of which was able to achieve strong performance. This hints at a phenomenon that will become clearer in the next section (which discusses the ANLI benchmark) – GPT-3 appears to be weak in the few-shot or one-shot setting at some tasks that involve comparing two sentences or snippets, for example whether a word is used the same way in two sentences (WiC), whether one sentence is a paraphrase of another, or whether one sentence implies another.This could also explain the comparatively low scores for RTE and CB, which also follow this format. Despite these weaknesses, GPT-3 still outperforms a fine-tuned BERT-large on four of eight tasks and on two tasks GPT-3 is close to the state-of-the-art held by a fine-tuned 11 billion parameter model.

WiC上表现较差，few-shot情况下的表现为49.4%。我们尝试了多种不同的短语和公式来处理WiC（它涉及确定一个单词在两个句子中是否具有相同的含义），但没有一种能够取得较好的表现。这说明了一个现象，在下一节（讨论ANLI基准测试）中将变得更加清晰——GPT-3在某些涉及比较两个句子或片段的任务中，如一个单词在两个句子中是否以相同的方式使用（WiC），一个句子是否是另一个句子的释义，或一个句子是否暗示另一个句子，few-shot或one-shot情况下表现较弱。这也可以解释RTE和CB的相对较低分数，它们也遵循这种格式。尽管存在这些弱点，GPT-3仍然在八项任务中的四项中优于经过微调的BERT-Large，在两项任务中，GPT-3接近经过微调的110亿参数模型的SOTA。

Finally, we note that the few-shot SuperGLUE score steadily improves with both model size and with number of examples in the context showing increasing benefits from in-context learning (Figure 3.8). We scale K up to 32 examples per task, after which point additional examples will not reliably fit into our context. When sweeping over values of K, we find that GPT-3 requires less than eight total examples per task to outperform a fine-tuned BERT-Largeon overall SuperGLUE score.

最后，我们注意到少样本情况下的SuperGLUE得分随着模型大小和上下文中的示例数量而稳步提高，显示出上下文学习的不断增强的益处（图3.8）。我们将K扩展到每个任务32个示例，此后，额外的示例将无法可靠地适合我们的上下文。在扫描K的值时，我们发现GPT-3每个任务只需要不到8个示例就能在整体SuperGLUE得分上优于经过微调的BERT-Large。

图3.8 SuperGLUE上的表现随着模型大小和上下文数量提升而升

3.8 NLI

Natural Language Inference (NLI) [Fyo00] concerns the ability to understand the relationship between two sentences.In practice, this task is usually structured as a two or three class classification problem where the model classifies whether the second sentence logically follows from the first, contradicts the first sentence, or is possibly true (neutral).SuperGLUE includes an NLI dataset, RTE, which evaluates the binary version of the task. On RTE, only the largest version of GPT-3 performs convincingly better than random (56%) in any evaluation setting, but in a few-shot settingGPT-3 performs similarly to a single-task fine-tuned BERT Large. We also evaluate on the recently introduced Adversarial Natural Language Inference (ANLI) dataset [NWD+19]. ANLI is a difficult dataset employing a series of adversarially mined natural language inference questions in three rounds (R1, R2, and R3). Similar to RTE, all of our models smaller than GPT-3 perform at almost exactly random chance on ANLI, even in the few-shot setting (~ 33%),whereas GPT-3 itself shows signs of life on Round 3. Results for ANLI R3 are highlighted in Figure 3.9 and full resultsfor all rounds can be found in Appendix H. These results on both RTE and ANLI suggest that NLI is still a very difficult task for language models and they are only just beginning to show signs of progress.

自然语言推理（NLI）[Fyo00]是指理解两个句子之间的关系。在实践中，这个任务通常被构造成一个二分类或三分类问题，模型会判断第二个句子是否逻辑上从第一个句子中推导出来，是否与第一个句子相矛盾，或者是可能为真（中性）。SuperGLUE包括一个NLI数据集RTE，它评估了这个任务的二元版本。在RTE上，只有GPT-3的最大版本在任何评估环境中都表现出比随机（56%）更好的表现，但在少样本情况下，GPT-3的表现与单任务微调的BERT Large相似。我们还在最近推出的对抗自然语言推理（ANLI）数据集[NWD+19]上进行了评估。ANLI是一个困难的数据集，使用一系列对抗性挖掘的自然语言推理问题进行三轮（R1、R2和R3）评估。与RTE类似，我们所有比GPT-3小的模型在ANLI上的表现几乎都是随机的，即使在少样本情况下也是如此（约33%），而GPT-3本身在第3轮中表现出了生命迹象。ANLI R3的结果在图3.9中突出显示，所有轮次的完整结果可以在附录H中找到。这些RTE和ANLI的结果表明，NLI对于语言模型仍然是一个非常困难的任务，它们只是开始显示出进展的迹象。

图3.9 GPT-3在ANLI数据集上三轮的表现

3.9 合成和定性任务

One way to probe GPT-3’s range of abilities in the few-shot (or zero- and one-shot) setting is to give it tasks which require it to perform simple on-the-fly computational reasoning, recognize a novel pattern that is unlikely to have occurred in training, or adapt quickly to an

unusual task. We devise several tasks to test this class of abilities. First, we test GPT-3’s ability to perform arithmetic. Second, we create several tasks that involve rearranging or unscrambling the letters in a word, tasks which are unlikely to have been exactly seen during training. Third, we test GPT-3’s ability to solve SAT-style analogy problems few-shot. Finally, we test GPT-3 on several qualitative tasks, including using new words in a sentence, correcting English grammar, and news article generation. We will release the synthetic datasets with the hope of stimulating further study of test-time behavior of language models.
探究GPT-3在少样本（或零样本和一次样本）情况下的能力范围的一种方法是给它一些需要进行即时计算推理、识别训练中不太可能出现的新模式或快速适应不寻常任务的任务。我们设计了几个任务来测试这类能力。首先，我们测试了GPT-3进行算术运算的能力。其次，我们创建了几个涉及重新排列或解密单词中字母的任务，这些任务在训练中不太可能被完全看到。第三，我们测试了GPT-3在少样本情况下解决SAT风格的类比问题的能力。最后，我们在几个定性任务上测试了GPT-3，包括在句子中使用新单词、纠正英语语法和新闻文章生成。我们将发布这些合成数据集，希望能够激发对语言模型测试时间行为的进一步研究。

3.9.1 算术

To test GPT-3’s ability to perform simple arithmetic operations without task-specific training, we developed a small battery of 10 tests that involve asking GPT-3 a simple arithmetic problem in natural language:

• 2 digit addition (2D+) – The model is asked to add two integers sampled uniformly from [0, 100), phrased in the form of a question, e.g. “Q: What is 48 plus 76? A: 124.”

• 2 digit subtraction (2D-) – The model is asked to subtract two integers sampled uniformly from [0, 100); the answer may be negative. Example: “Q: What is 34 minus 53? A: -19”.

• 3 digit addition (3D+) – Same as 2 digit addition, except numbers are uniformly sampled from [0, 1000).

• 3 digit subtraction (3D-) – Same as 2 digit subtraction, except numbers are uniformly sampled from [0, 1000).• 4 digit addition (4D+) – Same as 3 digit addition, except uniformly sampled from [0, 10000).

• 4 digit subtraction (4D-) – Same as 3 digit subtraction, except uniformly sampled from [0, 10000).• 5 digit addition (5D+) – Same as 3 digit addition, except uniformly sampled from [0, 100000).

• 5 digit subtraction (5D-) – Same as 3 digit subtraction, except uniformly sampled from [0, 100000).

• 2 digit multiplication (2Dx) – The model is asked to multiply two integers sampled uniformly from [0, 100),e.g. “Q: What is 24 times 42? A: 1008”.

• One-digit composite (1DC) – The model is asked to perform a composite operation on three 1 digit numbers,with parentheses around the last two. For example, “Q: What is 6+

(4*8)? A: 38”. The three 1 digit numbersare selected uniformly on [0, 10) and the operations are selected uniformly from {+,-,*}.
In all 10 tasks the model must generate the correct answer exactly. For each task we generate a dataset of 2,000 random instances of the task and evaluate all models on those instances.
为了测试GPT-3在没有特定任务训练的情况下执行简单算术运算的能力，我们开发了一个小型测试集，其中包括10个测试，涉及用自然语言问GPT-3一个简单的算术问题：
• 两位数加法（2D+）-模型被要求将两个从[0,100)中均匀抽样的整数相加，以问题的形式表达，例如“Q：48加76是多少？A：124。”
• 两位数减法（2D-）-模型被要求从[0,100)中均匀抽样两个整数相减；答案可能为负数。例如：“Q：34减53是多少？A：-19”。
• 三位数加法（3D+）-与两位数加法相同，只是数字从[0,1000)中均匀抽样。 • 三位数减法（3D-）-与两位数减法相同，只是数字从[0,1000)中均匀抽样。
• 四位数加法（4D+）-与三位数加法相同，只是从[0,10000)中均匀抽样。 • 四位数减法（4D-）-与三位数减法相同，只是从[0,10000)中均匀抽样。
• 五位数加法（5D+）-与三位数加法相同，只是从[0,100000)中均匀抽样。
• 五位数减法（5D-）-与三位数减法相同，只是从[0,100000)中均匀抽样。
• 两位数乘法（2Dx）-模型被要求将两个从[0,100)中均匀抽样的整数相乘，例如“Q：24乘以42是多少？A：1008”。
• 一位数字复合（1DC）-模型被要求对三个1位数字执行复合操作，最后两个数字用括号括起来。例如，“Q：6 +（4 * 8）是多少？A：38”。三个1位数字在[0,10)上均匀选择，操作从{+，-，*}中均匀选择。
在所有10个任务中，模型必须完全生成正确答案。对于每个任务，我们生成一个包含2,000个随机实例的数据集，并在这些实例上评估所有模型。
First we evaluate GPT-3 in the few-shot setting, for which results are shown in Figure 3.10. On addition and subtraction,GPT-3 displays strong proficiency when the number of digits is small, achieving 100% accuracy on 2 digit addition,98.9% at 2 digit subtraction, 80.2% at 3 digit addition, and 94.2% at 3-digit subtraction. Performance decreases as the number of digits increases, but GPT-3 still achieves 25-26% accuracy on four digit operations and 9-10% accuracy on five digit operations, suggesting at least some capacity to generalize to larger numbers of digits. GPT-3 also achieves 29.2% accuracy at 2 digit multiplication, an especially computationally intensive operation. Finally, GPT-3 achieves 21.3% accuracy at single digit combined operations (for example, 9*(7+5)), suggesting that it has some robustness beyond just single operations.
首先，我们在少样本情况下评估了GPT-3，其结果如图3.10所示。在加法和减法方面，当数字位数较小时，GPT-3表现出很强的熟练度，在2位数加法上达到100%的准确率，在2位数减法上达到98.9%的准确率，在3位数加法上达到80.2%的准确率，在3位数减法上达到94.2%的准确率。随着数字位数的增加，性能会下降，但GPT-3仍然在四位数运算上达到25-26%的准确率，在五位数运算上达到9-10%

的准确率，这表明它至少具有一定的推广到更大数字位数的能力。GPT-3在2位数乘法上也达到了29.2%的准确率，这是一项特别需要计算的操作。最后，GPT-3在单个数字组合操作（例如，9 *（7 + 5））上达到了21.3%的准确率，这表明它具有一些超越单个操作的鲁棒性。

图3.10 few-shot设置下不同大小模型在10个算术任务上的表现

As Figure 3.10 makes clear, small models do poorly on all of these tasks – even the 13 billion parameter model (the second largest after the 175 billion full GPT-3) can solve 2 digit addition and subtraction only half the time, and all other operations less than 10% of the time.

正如图3.10所示，小型模型在所有这些任务上表现不佳——即使是130亿参数模型（仅次于1750亿的完整GPT-3），也只能在2位数加减法中解决一半的问题，其他所有操作的准确率都不到10%。

One-shot and zero-shot performance are somewhat degraded relative to few-shot performance, suggesting that adaptation to the task (or at the very least recognition of the task) is important to performing these computations correctly.Nevertheless, one-shot performance is still quite strong, and even zero-shot performance of the full GPT-3 significantly outperforms few-shot learning for all smaller models. All three settings for the full GPT-3 are shown in Table 3.9, and model capacity scaling for all three settings is shown in Appendix H.

相对于few-shot情况下的表现，ont-shot和zero-shot情况下的表现有所下降，这表明适应任务（或至少识别任务）对于正确执行这些计算很重要。尽管如此，one-shot的表现仍然相当强，甚至完整的GPT-3的zero-shot表现也显著优于所有较小模型的few-shot。完整的GPT-3的所有三个设置都显示在表3.9中，所有三个设置的模型容量缩放都显示在附录H中。

表3.9 GPT-3 1750亿参数，不同设置在基本算术任务上的结果

To spot-check whether the model is simply memorizing specific arithmetic problems, we took the 3-digit arithmetic problems in our test set and searched for them in our training data in both the forms " + =" and" plus ". Out of 2,000 addition problems we found only 17 matches (0.8%) and out of 2,000 subtraction problems we found only 2 matches (0.1%), suggesting that only a trivial fraction of the correct answers could have been memorized. In addition, inspection of incorrect answers reveals that the model often makes mistakes such as not carrying a “1”, suggesting it is actually attempting to perform the relevant computation rather thanmemorizing a table.

为了检查模型是否只是记忆了特定的算术问题，我们在测试集中选取了3位数的算术问题，并在训练数据中搜索了这些问题，包括“ + =”和“ plus ”两种形式。在2000个加法问题中，我们只找到了17个匹配项（0.8%），在2000个减法问题中，我们只找到了2个匹配项（0.1%），这表明只有微不足道的正确答案可能已经被记忆。此外，对错误答案的检查表明，模型经常犯错误，例如没有携带“1”，这表明它实际上正在尝试执行相关的计算，而不是记忆表格。

Overall, GPT-3 displays reasonable proficiency at moderately complex arithmetic in few-shot, one-shot, and even zero-shot settings.
总的来说，GPT-3在few-shot、one-shot和zero-shot设置下表现出了合理的能力，能够处理中等复杂的算术问题。
3.9.2 单词混淆和操作任务
To test GPT-3’s ability to learn novel symbolic manipulations from a few examples, we designed a small battery of 5 “character manipulation” tasks. Each task involves giving the model a word distorted by some combination of scrambling, addition, or deletion of characters, and asking it to recover the original word. The 5 tasks are:
• Cycle letters in word (CL) – The model is given a word with its letters cycled, then the “=” symbol, and is expected to generate the original word. For example, it might be given “lyinevitab” and should output “inevitably”.
• Anagrams of all but first and last characters (A1) – The model is given a word where every letter except the first and last have been scrambled randomly, and must output the original word. Example: criroptuon =corruption.
• Anagrams of all but first and last 2 characters (A2) – The model is given a word where every letter except the first 2 and last 2 have been scrambled randomly, and must recover the original word. Example: opoepnnt→ opponent.
• Random insertion in word (RI) – A random punctuation or space character is inserted between each letter of a word, and the model must output the original word. Example: s.u!c/c!e.s s i/o/n = succession.
• Reversed words (RW) – The model is given a word spelled backwards, and must output the original word.Example: stcejbo → objects.
为了测试GPT-3从少量示例中学习新的符号操作的能力，我们设计了一个由5个“字符操作”任务组成的小型测试集。每个任务都涉及对单词进行一些混淆和操作，例如字母混淆、添加或删除，然后要求模型恢复原始单词。这5个任务分别是：
• 单词字母循环（CL）- 模型会得到一个字母循环的单词，然后是“=”符号，期望输出原始单词。例如，它可能会得到“lyinevitab”，并应输出“inevitably”。
• 除第一个和最后一个字符外的所有字符的字谜（A1）- 模型会得到一个单词，其中除第一个和最后一个字符外的每个字母都被随机混淆，必须输出原始单词。例如：criroptuon = corruption。
• 除前两个和最后两个字符外的所有字符的字谜（A2）- 模型会得到一个单词，其中除前两个和后两个字符外的每个字母都被随机混淆，必须恢复原始单词。例如：opoepnnt→opponent。
• 单词中的随机插入（RI）- 在单词的每个字母之间插入随机标点符号或空格，模型必须输出原始单词。例如：s.u!c/c!e.s s i/o/n = succession。
• 单词反转（RW）- 模型会得到一个单词的反向拼写，并必须输出原始单词。例如：stcejbo→objects。

For each task we generate 10,000 examples, which we chose to be the top 10,000 most frequent words as measured by[Nor09] of length more than 4 characters and less than 15 characters. The few-shot results are shown in Figure 3.11.Task performance tends to grow smoothly with model size, with the full GPT-3 model achieving 66.9% on removing random insertions, 38.6% on cycling letters, 40.2% on the easier anagram task, and 15.1% on the more difficult anagram task (where only the first and last letters are held fixed). None of the models can reverse the letters in a word.
对于每个任务，我们生成了10,000个示例，这些示例是由[Nor09]测量的长度大于4个字符且小于15个字符的前10,000个最常见的单词。图3.11显示了少量样本的结果。随着模型大小的增加，任务性能往往会平稳增长，完整的GPT-3模型在去除随机插入方面的准确率为66.9%，在字母循环方面的准确率为38.6%，在更容易的字谜任务方面的准确率为40.2%，在更困难的字谜任务方面的准确率为15.1%（只有第一个和最后一个字母被固定）。没有一个模型能够颠倒单词中的字母。

表3.10 GPT-3 1750亿参数在不同的单词混淆操作任务上的表现

In the one-shot setting, performance is significantly weaker (dropping by half or more), and in the zero-shot setting the model can rarely perform any of the tasks (Table 3.10). This suggests that the model really does appear to learn these tasks at test time, as the model cannot perform them zero-shot and their artificial nature makes them unlikely to appear in the pre-training data (although we cannot confirm this with certainty).
在one-shot设置中，性能较差（下降一半或更多），在zero-shot设置中，模型很少能够执行任何任务（表3.10）。这表明，模型确实似乎在测试时学习了这些任务，它们的人工生成的性质使它们不太可能出现在预训练数据中（尽管我们无法确定这一点）。
We can further quantify performance by plotting “in-context learning curves”, which show task performance as a function of the number of in-context examples. We show in-context learning curves for the Symbol Insertion taskin Figure 1.2. We can see that larger models are able to make increasingly effective use of in-context information,including both task examples and natural language task descriptions.
我们可以通过绘制“上下文学习曲线”来进一步量化性能，这些曲线显示了任务性能作为上下文示例数量的函数。我们在图1.2中展示了符号插入任务的上下文学习曲线。我们可以看到，更大的模型能够越来越有效地利用上下文信息，包括任务示例和自然语言任务描述。
Finally, it is worth adding that solving these tasks requires character-level manipulations, whereas our BPE encoding operates on significant fractions of a word (on average ~ 0.7 words per token), so from the LM’s perspective succeeding at these tasks involves not just manipulating BPE tokens but understanding and pulling apart their substructure. Also,CL, A1, and A2 are not bijective (that is, the unscrambled word is not a deterministic function of the scrambled word),requiring the model to perform some search to find the correct

unscrambling. Thus, the skills involved appear to requirenon-trivial pattern-matching and computation.
最后，值得补充的是，解决这些任务需要对字符进行操作，而我们的BPE编码操作的是单词的重要部分（平均每个标记约为0.7个单词），因此从LM的角度来看，成功完成这些任务不仅涉及操作BPE标记，还涉及理解和分解它们的子结构。此外，CL、A1和A2不是双射的（即未混淆的单词不是混淆的单词的确定性函数），需要模型执行一些搜索才能找到正确的解密。因此，所涉及的技能似乎需要进行复杂的模式匹配和计算。

图3.11 不同模型大小的few-shot在五个单词混淆任务的表现

3.9.3 SAT类比题

To test GPT-3 on another task that is somewhat unusual relative to the typical distribution of text, we collected a set of 374 "SAT analogy" problems [TLBS03]. Analogies are a style of multiple choice question that constituted a section of the SAT college entrance exam before 2005. A typical example is "audacious is to boldness as (a) sanctimonious is tohypocrisy, (b) anonymous is to identity, (c) remorseful is to misdeed, (d) deleterious is to result, (e) impressionable is totemptation". The student is expected to choose which of the five word pairs has the same relationship as the original word pair; in this example the answer is "sanctimonious is to hypocrisy". On this task GPT-3 achieves 65.2% in the few-shot setting, 59.1% in the one-shot setting, and 53.7% in the zero-shot setting, whereas the average score among college applicants was 57% [TL05] (random guessing yields 20%). As shown in Figure 3.12, the results improve withscale, with the the full 175 billion model improving by over 10% compared to the 13 billion parameter model.

图3.12 不同模型尺寸在zero-shot、one-shot和few-shot上进行SAT类比题的任务表现

为了测试GPT-3在相对于典型文本分布而言有些不寻常的任务上的表现，我们收集了一组374个"SAT类比"问题[TLBS03]。类比是一种多项选择题，它在2005年之前是SAT大学入学考试的一部分。一个典型的例子是"audacious is to boldness as (a) sanctimonious is tohypocrisy, (b) anonymous is to identity, © remorseful is to misdeed, (d) deleterious is to result, (e) impressionable is totemptation"。学生需要选择哪个五个单词对与原始单词对具有相同的关系；在这个例子中，答案是"sanctimonious is to hypocrisy"。在这个任务中，GPT-3在few-shot的情况下的准确率为65.2%，在one-shot的情况下的准确率为59.1%，在zero-shot的情况下的准确率为53.7%，而大学申请者的平均分数为57%[TL05]（随机猜测的准确率为20%）。如图3.12所示，随着规模的扩大，效果逐渐提升，全1750亿模型的准确率比13亿参数模型提高了10%以上。

3.9.4 新闻文章生成

Previous work on generative language models qualitatively tested their ability to generate synthetic "news articles" by conditional sampling from the model given a human-written prompt consisting of a plausible first sentence for a news story [RWC+19]. Relative to

[RWC+19], the dataset used to train GPT-3 is much less weighted towards news articles,so trying to generate news articles via raw unconditional samples is less effective – for example GPT-3 often interprets the proposed first sentence of a “news article” as a tweet and then posts synthetic responses or follow-up tweets. To solve this problem we employed GPT-3’s few-shot learning abilities by providing three previous news articles in the model’s context to condition it. With the title and subtitle of a proposed next article, the model is able to reliably generate short articles in the “news” genre.
以往的生成式语言模型定性地测试了合成的“新闻文章”的能力[RWC+19]，它们通过人类编写的提示来生成。相对于[RWC+19]，用于训练GPT-3的数据集更少地偏向于新闻文章，因此尝试通过原始无条件样本生成新闻文章的效果较差-例如，GPT-3经常将“新闻文章”的建议第一句解释为推文，然后发布合成响应或后续推文。为了解决这个问题，我们利用了GPT-3的少量样本学习能力，通过提供三篇先前的新闻文章来对其进行调整。在提议的下一篇文章的标题和副标题的情况下，该模型能够可靠地生成“新闻”类型的短文章。
To gauge the quality of news article generation from GPT-3 (which we believe is likely to be correlated with conditional sample generation quality in general), we decided to measure human ability to distinguish GPT-3-generated articles from real ones. Similar work has been carried out by Kreps et al. [KMB20] and Zellers et al. [ZHR+19]. Generative language models are trained to match the distribution of content generated by humans, so the (in)ability of humans to distinguish the two is a potentially important measure of quality.
为了衡量GPT-3生成新闻文章的质量，我们决定测量人类区分GPT-3生成的文章和真实文章的能力。Kreps等人[KMB20]和Zellers等人[ZHR+19]也进行了类似的工作。生成式语言模型的训练目标是匹配人类生成的内容分布，因此人类区分两者的能力（或无能力）是一个潜在的重要质量指标。
In order to see how well humans can detect model generated text, we arbitrarily selected 25 article titles and subtitles from the website newser.com (mean length: 215 words). We then generated completions of these titles and subtitlesfrom four language models ranging in size from 125M to 175B (GPT-3) parameters (mean length: 200 words). For each model, we presented around 80 US-based participants with a quiz consisting of these real titles and subtitles followedby either the human written article or the article generated by the model. Participants were asked to select whether the article was “very likely written by a human”, “more likely written by a human”, “I don’t know”, “more likely written bya machine”, or “very likely written by a machine”.
为了测试人类区分模型生成文本的能力，我们随机选择了25个文章标题和副标题，这些标题和副标题来自newser.com网站（平均长度：215个单词）。然后，我们从四个语言模型中生成了这些标题和副标题的完成情况，这些模型的大小从125M到175B（GPT-3）参数不等（平均长度：200个单词）。对于每个模型，我们向大约80名美国参与者提供了一个测验，包括这些真实的标题和副标题，然后是人类编写的文章或模型生成的文章。参与者被要求选择文章是“非常可能由人类编写”，“更可能由人类编写”，“我不知道”，“更可能由机器编写”还是“非常可能由机器编写”。

The articles we selected were not in the models' training data and the model outputs were formatted and selected programmatically to prevent human cherry-picking. All models used the same context to condition outputs on and were pre-trained with the same context size and the same article titles and subtitles were used as prompts for each model.However, we also ran an experiment to control for participant effort and attention that followed the same format but involved intentionally bad model generated articles. This was done by generating articles from a "control model": a160M parameter model with no context and increased output randomness.

我们选择的文章不在模型的训练数据中，模型输出是经过格式化和程序化选择的，以防止人为地挑选。所有模型都使用相同的上下文来调整输出，并使用相同的上下文大小进行预训练，每个模型的提示都使用相同的文章标题和副标题。但是，我们还进行了一个实验，以控制参与者的努力和注意力，该实验遵循了相同的格式，但故意生成的错误模型生成的文章。这是通过从"控制模型"生成文章来完成的：一个没有上下文且输出随机性增加的160M参数模型。

Mean human accuracy (the ratio of correct assignments to non-neutral assignments per participant) at detecting that the intentionally bad articles were model generated was ~ 86% where 50% is chance level performance. By contrast,mean human accuracy at detecting articles that were produced by the 175B parameter model was barely above chance at ~ 52% (see Table 3.11).Human abilities to detect model generated text appear to decrease as model size increases:there appears to be a trend towards chance accuracy with model size, and human detection of GPT-3 is close to chance.This is true despite the fact that participants spend more time on each output as model size increases (see Appendix E).

检测到有意制造的错误文章是模型生成的平均人类准确率（每个参与者的正确分配与非中性分配的比率）约为86%，其中50%是机会水平表现。相比之下，检测到由175B参数模型生成的文章的平均人类准确率仅略高于机会水平，约为52%（见表3.11）。随着模型大小的增加，人类检测模型生成的文本的能力似乎会降低：随着模型大小的增加，出现了趋向于机会准确性的趋势，而GPT-3的人类检测接近机会水平。尽管参与者在模型大小增加时会花费更多时间在每个输出上（请参见附录E），但这是真实的。

表3.11 检测短新闻文章是否由模型生成的人类准确率

图3.13 人类区分新闻文章是否由模型生成

表3.12 人类区分500字文章是否由模型生成

Examples of synthetic articles from GPT-3 are given in Figures 3.14 and 3.15.7 Much of the text is—as indicated by theevaluations—difficult for humans to distinguish from authentic human content. Factual inaccuracies can be an indicatorthat an article is model generated since, unlike human authors, the models have no access to the specific facts that the article titles refer to or when the article was written. Other indicators include repetition, non

sequiturs, and unusual phrasings, though these are often subtle enough that they are not noticed.
GPT-3合成文章的示例如图3.14和3.15所示。根据评估，其中大部分文本很难与真实的人类内容区分开来。事实上不准确可能是文章是模型生成的指标，因为与人类作者不同，模型无法访问文章标题所指的具体事实或文章编写时间。其他指标包括重复、不合逻辑的推论和不寻常的措辞，尽管这些指标通常足够微妙，以至于人们不会注意到。

图3.14 GPT-3生成的新闻文章，人类很区分

3.15 GPT-3生成的新闻文章，人类很区分

Related work on language model detection by Ippolito et al. [IDCBE19] indicates that automatic discriminators likeG R O V E R [ZHR+19] and GLTR [GSR19] may have greater success at detecting model generated text than humanevaluators. Automatic detection of these models may be a promising area of future research.
由Ippolito等人[IDCBE19]进行的有关语言模型检测的相关工作表明，像GROVER[ZHR+19]和GLTR[GSR19]这样的自动鉴别器可能比人类评估者更成功地检测到模型生成的文本。自动检测这些模型可能是未来研究的一个有前途的领域。
Ippolito et al. [IDCBE19] also note that human accuracy at detecting model generated text increases as humans observemore tokens. To do a preliminary investigation of how good humans are at detecting longer news articles generatedby GPT-3 175B, we selected 12 world news articles from Reuters with an average length of 569 words and generatedcompletions of these articles from GPT-3 with an average length of 498 words (298 words longer than our initialexperiments). Following the methodology above, we ran two experiments, each on around 80 US-based participants, tocompare human abilities to detect the articles generated by GPT-3 and a control model.
Ippolito等人[IDCBE19]还指出，随着人类观察更多的标记，人类检测模型生成的文本的准确性会提高。为了初步调查人类在检测由GPT-3 175B生成的较长新闻文章方面的能力，我们从路透社选择了12篇平均长度为569个单词的世界新闻文章，并从GPT-3中生成了这些文章的完成情况，平均长度为498个单词（比我们最初的实验长298个单词）。按照上述方法，我们进行了两个实验，每个实验约有80名美国参与者，以比较人类检测GPT-3生成的文章和控制模型生成的文章的能力。
We found that mean human accuracy at detecting the intentionally bad longer articles from the control model was~ 88%, while mean human accuracy at detecting the longer articles that were produced by GPT-3 175B was still barelyabove chance at ~ 52% (see Table 3.12). This indicates that, for news articles that are around 500 words long, GPT-3continues to produce articles that humans find difficult to distinguish from human written news articles.
我们发现，人类平均准确率在检测控制模型中故意制造的较长错误文章时约为88%，而在检测由GPT-3 175B生成的较长文章时，人类平均准确率仅略高于机会水平，约为52%（见表3.12）。这表明，对于大约500个单词的新闻文章，GPT-3继续生成人类难以区分的新闻文章。

3.9.6 纠正英语语法

Another task well suited for few-shot learning is correcting English grammar. We test this with GPT-3 in the few-shot setting by giving prompts of the form "Poor English Input: \n Good English Output:<sentence>". We give GPT-3 one human-generated correction and then ask it to correct 5 more (again without anyomissions or repeats). Results are shown in Figure 3.17.

另一个适合少量样本学习的任务是纠正英语语法。我们使用GPT-3在few-shot样本的情况下进行测试，通过给出“Poor English Input: \n Good English Output:<sentence>”的提示。我们给GPT-3一个人类生成的更正，然后要求它纠正5个（再次没有任何省略或重复）。结果如图3.17所示。

3.17 few-shot GPT-3纠正英语语法

4 检测和避免基线记忆

Since our training dataset is sourced from the internet, it is possible that our model was trained on some of our benchmark test sets. Accurately detecting test contamination from internet-scale datasets is a new area of research without established best practices. While it is common practice to train large models without investigating contamination,given the increasing scale of pretraining datasets, we believe this issue is becoming increasingly important to attend to.

由于我们的训练数据集来自互联网，因此我们的模型有可能在我们的基准测试集中进行了训练。准确检测互联网规模的数据集中的测试污染是一个新的研究领域，目前还没有确定的最好的办法。虽然通常的做法是在不调查污染的情况下训练大型模型，但考虑到预训练数据集的规模不断增加，我们认为这个问题变得越来越重要。

This concern is not just hypothetical. One of the first papers to train a language model on Common Crawl data [TL18]detected and removed a training document which overlapped with one of their evaluation datasets. Other work such as GPT-2 [RWC+19] also conducted post-hoc overlap analysis. Their study was relatively encouraging, finding that although models did perform moderately better on data that overlapped between training and testing, this did not significantly impact reported results due to the small fraction of data which was contaminated (often only a few percent).

这个问题不仅是假设的，业内已有研究。第一篇在Common Crawl数据上训练语言模型的论文之一[TL18]检测到并删除了一个与他们的评估数据集重叠的训练文档。其他工作，如GPT-2 [RWC+19]也进行了事后重叠分析。他们的研究相对鼓舞人心，发现虽然模型在训练和测试之间重叠的数据上表现得更好，但由于受污染的数据只占很小一部分，这并没有对报告的结果产生显著影响（通常只有几个百分点）。

GPT-3 operates in a somewhat different regime. On the one hand, the dataset and model size are about two orders of magnitude larger than those used for GPT-2, and include a large amount of Common Crawl, creating increased potential for contamination and memorization.

On the other hand, precisely due to the large amount of data, even GPT-3 175B does not overfit its training set by a significant amount, measured relative to a held-out validation set with which it was deduplicated (Figure 4.1). Thus, we expect that contamination is likely to be frequent, but that its effects may not be as large as feared.
GPT-3的运行方式与GPT-2略有不同。一方面，数据集和模型大小约为GPT-2使用的数据集和模型大小的两个数量级，并包括大量的Common Crawl，从而增加了污染和记忆的潜力。另一方面，正是由于数据量很大，即使是GPT-3 175B也没有过度拟合其训练集，相对于它被去重的保留验证集（图4.1），过度拟合的程度并不显著。因此，我们预计污染可能会经常发生，但其影响可能不会像人们担心的那样大。
We initially tried to address the issue of contamination by proactively searching for and attempting to remove any over lapbetween our training data and the development and test sets of all benchmarks studied in this paper. Unfortunately, a bug resulted in only partial removal of all detected overlaps from the training data. Due to the cost of training, it wasn’t feasible to retrain the model. To address this, we investigate in detail how the remaining detected overlap impacts results.
我们最初试图通过主动搜索并尝试删除我们的训练数据与本文研究的所有基准测试集的开发和测试集之间的任何重叠来解决污染问题。不幸的是，一个错误导致只有部分重叠被从训练数据中删除。由于训练成本高昂，重新训练模型是不可行的。为了解决这个问题，我们详细研究了剩余检测到的重叠对结果的影响。
For each benchmark, we produce a ‘clean’ version which removes all potentially leaked examples, defined roughly as examples that have a 13-gram overlap with anything in the pretraining set (or that overlap with the whole example when it is shorter than 13-grams). The goal is to very conservatively flag anything that could potentially be contamination,so as to produce a clean subset that is free of contamination with high confidence. The exact procedure is detailed inAppendix C.
每个基准测试都会生成一个“干净”版本，该版本会删除所有可能泄露的示例，这些示例与预训练集中的任何内容重叠13个单词（或者当整个示例短于13个词时与整个示例重叠）。目标是非常保守地标记任何可能污染的内容，以便产生一个高度自信的无污染子集。详细的过程在附录C中有详细说明。
We then evaluate GPT-3 on these clean benchmarks, and compare to the original score. If the score on the clean subset is similar to the score on the entire dataset, this suggests that contamination, even if present, does not have a significant effect on reported results. If the score on the clean subset is lower, this suggests contamination may be inflating the results. The results are summarized in Figure 4.2. Although potential contamination is often high (with a quarter of benchmarks scoring over 50%), in most cases performance changes only negligibly, and we see no evidence that contamination level and performance difference are correlated. We conclude that either our conservative method substantially over estimated contamination or that contamination has little effect on performance.

我们随后在这些干净的基准测试上评估了GPT-3，并将其与原始分数进行了比较。如果清洁子集上的分数与整个数据集上的分数相似，则表明即使存在污染，其对报告的结果也没有显着影响。如果清洁子集上的分数较低，则表明污染可能会夸大结果。结果总结在图4.2中。尽管潜在污染通常很高（四分之一的基准测试得分超过50%），但在大多数情况下，性能变化仅微不足道，我们没有看到污染水平和性能差异之间的相关性证据。我们得出结论，要么我们的保守方法大大高估了污染，要么污染对性能影响不大。

## 4.2 基准测试污染分析

Below, we review in more detail the few specific cases where either (1) the model performs significantly worse on the cleaned version, or (2) potential contamination is very high, which makes measuring the performance difference difficult.

下面，我们将更详细地审查几种特定情况，其中（1）模型在清理后的版本上表现较差，或者（2）污染非常高，这使得测量性能差异变得困难。

Our analysis flagged six groups of benchmarks for further investigation: Word Scrambling, Reading Comprehension(QuAC, SQuAD2, DROP), PIQA, Winograd, language modeling tasks (Wikitext tasks, 1BW), and German to English translation. Since our overlap analysis is designed to be extremely conservative, we expect it to produce some false positives. We summarize the results for each group of tasks below:

我们的分析标记了六组基准测试以进行进一步调查：单词混淆，阅读理解（QuAC，SQuAD2，DROP），PIQA，Winograd，语言建模任务（Wikitext 任务，1BW）和德语到英语的翻译。由于我们的重叠分析非常保守，因此我们预计它会产生一些误报。我们在下面总结了每组任务的结果。

• Reading Comprehension: Our initial analysis flagged >90% of task examples from QuAC, SQuAD2, and DROP as potentially contaminated, so large that even measuring the differential on a clean subset was difficult.Upon manual inspection, however, we found that for every overlap we inspected, in all 3 datasets, the source text was present in our training data but the question/answer pairs were not, meaning the model gains only background information and can not memorize the answer to a specific question.

• 阅读理解：我们的初始分析标记了QuAC、SQuAD2和DROP中超过90%的示例，这些数据集非常大。在手动检查中，我们发现在我们检查的每个重叠部分中，所有3个数据集中的源文都存在于我们的训练数据中，但问题/答案对却不在其中，这意味着模型只获得了背景信息，无法记忆特定问题的答案。

• German translation: We found 25% of the examples in the WMT16 German-English test set were markedas potentially contaminated, with an associated total effect size of 1-2 BLEU. Upon inspection, none of the flagged examples contain paired sentences resembling NMT training data and collisions were monolingual matches mostly of snippets of events discussed in the news.

• 德语翻译：我们发现WMT16德英测试集中25%的示例可能存在污染，其相关的总效应大小为1-2 BLEU。经过检查，没有一个被标记的示例包含类似于NMT训练数据的成对句子，而冲突大多是新闻中讨论的片段的单语匹配。
• Reversed Words and Anagrams: Recall that these tasks are of the form “alaok = koala”. Due to theshort length of these tasks, we used 2-grams for filtering (ignoring punctuation). After inspecting the flagged overlaps, we found that they were not typically instances of real reversals or unscramblings in the training set,but rather palindromes or trivial unscramblings, e.g “kayak = kayak”. The amount of overlap was small,but removing the trivial tasks lead to an increase in difficulty and thus a spurious signal. Related to this, thesymbol insertion task shows high overlap but no effect on performance – this is because that task involvesremoving non-letter characters from a word, and the overlap analysis itself ignores such characters, leading tomany spurious matches.
• 反转单词和字谜：请回忆一下，这些任务的形式为“alaok = koala”。由于这些任务的长度很短，我们使用2-gram进行过滤（忽略标点符号）。在检查重叠部分后，我们发现它们通常不是训练集中真正的解密实例，而是回文或琐碎的解密，例如“kayak = kayak”。重叠量很小，但删除琐碎的任务会增加难度，从而产生虚假信号。同时，符号插入任务显示高重叠但对性能没有影响——这是因为该任务涉及从单词中删除非字母字符，而重叠分析本身忽略了这些字符，导致许多虚假的匹配。
• PIQA: The overlap analysis flagged 29% of examples as contaminated, and observed a 3 percent age point absolute decrease (4% relative decrease) in performance on the clean subset. Though the test dataset was released after our training set was created and its labels are hidden, some of the web pages used by thecrowdsourced dataset creators are contained in our training set. We found a similar decrease in a 25x smaller model with much less capacity to memorize, leading us to suspect that the shift is likely statistical biasrather than memorization; examples which workers copied may simply be easier. Unfortunately, we cannotrigorously prove this hypothesis. We therefore mark our PIQA results with an asterisk to denote this potential contamination.
PIQA: 重叠分析标记了29%的示例可能存在污染，并观察到干净子集上的性能绝对下降了3个百分点（相对下降了4%）。尽管测试数据集是在我们创建训练集之后发布的，其标签是隐藏的，但一些众包数据集创建者使用的网页包含在我们的训练集中。我们在一个25倍小的模型中发现了类似的下降，其记忆能力要小得多，这使我们怀疑这种转变很可能是统计偏差而不是记忆；人工复制的示例可能只是更容易。不幸的是，我们无法严格证明这个假设。因此，我们用星号标记我们的PIQA结果，以表示这种潜在污染。
• Winograd: The overlap analysis flagged 45% of examples, and found a 2.6% decrease in performance on the clean subset. Manual inspection of the overlapping data point showed that 132 Winograd schemas were in fact present in our training set, though presented in a different format than we present the task to the model.Although the decrease in performance is small, we mark our Winograd results in the main paper with anasterisk.

Winograd: 重叠分析标记了45%的示例可能存在污染，并发现干净子集上的性能绝对下降了2.6%。手动检查重叠数据点显示，132个Winograd模式实际上存在于我们的训练集中，但格式与任务中的不同。尽管性能下降很小，但我们在主要论文中用星号标记了我们的Winograd结果。

• Language modeling: We found the 4 Wikipedia language modeling benchmarks measured in GPT-2, plus theChildren’s Book Test dataset, to be almost entirely contained in our training data. Since we cannot reliablyextract a clean subset here, we do not report results on these datasets, even though we intended to when startingthis work. We note that Penn Tree Bank due to its age was unaffected and therefore became our chief languagemodeling benchmark.

语言建模：我们发现GPT-2测量的4个维基百科语言建模基准测试以及儿童读物测试数据集几乎全部包含在我们的训练数据中。由于我们无法在此处可靠地提取干净的子集，因此我们不会报告这些数据集的结果，即使我们在开始这项工作时打算这样做。我们注意到，由于其年代久远，Penn Tree Bank并未受到影响，因此成为我们的主要语言建模基准测试。

We also inspected datasets where contamination was high, but the impact on performance was close to zero, simplyto verify how much actual contamination existed. These appeared to often contain false positives. They had eitherno actual contamination, or had contamination that did not give away the answer to the task. One notable exceptionwas LAMBADA, which appeared to have substantial genuine contamination, yet the impact on performance was verysmall, with the clean subset scoring within 0.5% of the full dataset. Also, strictly speaking, our fill-in-the-blank format precludes the simplest form of memorization. Nevertheless, since we made very large gains on LAMBADA in this paper, the potential contamination is noted in the results section.

我们还检查了污染率很高但对性能影响接近于零的数据集，以验证实际污染的程度。这些数据集通常包含误报。它们要么没有实际污染，要么有的污染并不会泄露任务的答案。一个值得注意的例外是LAMBADA，它似乎存在着实质性的污染，但对性能的影响非常小，干净子集的得分与完整数据集的得分相差不到0.5%。严格来说，我们的填空格式排除了最简单的记忆形式。尽管如此，由于我们在本文中在LAMBADA上取得了非常大的进展，因此潜在的污染在结果还是记录了下来。

An important limitation of our contamination analysis is that we cannot be sure that the clean subset is drawn from the same distribution as the original dataset. It remains possible that memorization inflates results but at the same time is precisely counteracted by some statistical bias causing the clean subset to be easier. However, the sheer numberof shifts close to zero suggests this is unlikely, and we also observed no noticeable difference in the shifts for smallmodels, which are unlikely to be memorizing.

我们的污染分析的一个重要限制是，我们无法确定干净子集是否来自与原始数据集相同的分布。记忆可能会增强结果，但同时又被一些统计偏差精确地抵消，从而使干净子集更容易。然而，接近零的转变数量表明这种情况不太可能发生，而且我们还观察到小型模型的转变没有明显的差异，这些模型不太可能进行记忆。”

Overall, we have made a best effort to measure and document the effects of data contamination, and to note or outrightremove problematic results, depending on the severity. Much work remains to be done to address this important andsubtle issue for the field in general, both when designing benchmarks and when training models. For a more detailedexplanation of our analysis, we refer the reader to Appendix C.

总的来说，我们已经尽最大努力测量和记录数据污染的影响，并根据严重程度记录或直接删除问题结果。在设计基准测试和训练模型时，仍有许多工作需要做，以解决这个重要而微妙的问题。有关我们分析的更详细说明，请参见附录C。

5 限制

GPT-3 and our analysis of it have a number of limitations. Below we describe some of these and suggest directions for future work.

GPT-3和我们对它的分析有一些限制。下面我们列出了一些，并对未来的工作提出了方向。

First, despite the strong quantitative and qualitative improvements of GPT-3, particularly compared to its directpredecessor GPT-2, it still has notable weaknesses in text synthesis and several NLP tasks. On text synthesis, although the overall quality is high, GPT-3 samples still sometimes repeat themselves semantically at the document level, start to lose coherence over sufficiently long passages, contradict themselves, and occasionally contain non-sequitur sentences or paragraphs. We will release a collection of 500 uncurated unconditional samples to help provide a better sense of GPT-3’s limitations and strengths at text synthesis. Within the domain of discrete language tasks, we have noticed informally that GPT-3 seems to have special difficulty with “common sense physics”, despite doing well on some datasets (such as PIQA [BZB+19]) that test this domain. Specifically GPT-3 has difficulty with questions of the type“If I put cheese into the fridge, will it melt?”. Quantitatively, GPT-3’s in-context learning performance has some notable gaps on our suite of benchmarks, as described in Section 3, and in particular it does little better than chance when evaluated one-shot or even few-shot on some “comparison” tasks, such as determining if two words are used the sameway in a sentence, or if one sentence implies another (WIC and ANLI respectively), as well as on a subset of reading comprehension tasks. This is especially striking given GPT-3’s strong few-shot performance on many other tasks.

首先，尽管GPT-3在定量和定性方面都有很大的改进，特别是与其直接前身GPT-2相比，但它在文本合成和几个NLP任务方面仍然存在明显的弱点。在文本合成方面，尽管总体质量很高，GPT-3样本有时仍会在文档级别上语义重复，长段落开始失去连贯性，自相矛盾，并偶尔包含不合逻辑的句子或段落。我们将发布500个未经筛选的无条件样本集，以帮助更好地了解GPT-3在文本合成方面的限制和优势。在离散语言任务领域，我们非正式地注意到，尽管在一些测试该领域的数据集（例如PIQA [BZB+19]）上表现良好，但GPT-3似乎在“常识物理”方面有特殊困难。具体而言，GPT-3在以下类型的问题上存在困难：“如果我把奶酪放进冰箱，它会融化吗？”。定量上，GPT-3在我们的基准测试套件中的上下文学习性能存在一些显著差距，如第3节所述，特别是在一些“比较”任务上，例如确

定两个单词在句子中的使用方式是否相同，或者一个句子是否暗示另一个句子（分别为WIC和ANLI），以及在阅读理解任务的子集上。考虑到GPT-3在许多其他任务上的强大的fwe-shot性能，这一点尤其引人注目。

GPT-3 has several structural and algorithmic limitations, which could account for some of the issues above. We focused on exploring in-context learning behavior in auto regressive language models because it is straightforward to both sample and compute likelihoods with this model class. As a result our experiments do not include any bidirectional architectures or other training objectives such as denoising. This is a noticeable difference from much of the recent literature, which has documented improved fine-tuning performance when using these approaches over standard language models [RSR+19]. Thus our design decision comes at the cost of potentially worse performance on tasks which empirically benefit from bidirectionality. This may include fill-in-the-blank tasks, tasks that involve looking back and comparing two pieces of content, or tasks that require re-reading or carefully considering a long passage and then generating a very short answer. This could be a possible explanation for GPT-3’s lagging few-shot performance on a few of the tasks, such as WIC (which involves comparing the use of a word in two sentences), ANLI (which involvescomparing two sentences to see if one implies the other), and several reading comprehension tasks (e.g. QuAC andRACE). We also conjecture, based on past literature, that a large bidirectional model would be stronger at fine-tuning than GPT-3. Making a bidirectional model at the scale of GPT-3, and/or trying to make bidirectional models work with few- or zero-shot learning, is a promising direction for future research, and could help achieve the “best of both worlds”.

GPT-3存在一些结构和算法上的限制，这可能解释了上述一些问题。我们专注于探索自回归语言模型的上下文学习行为，因为使用这种模型类别进行采样和计算似乎很简单。因此，我们的实验不包括任何双向架构或其他训练目标，例如去噪。最近一些文献通过使用这些方法获得了优于标准语言模型的效果。因此，我们的设计与双向性的任务相比性能更差。这可能包括填空任务、涉及回顾和比较两个内容的任务，或需要重新阅读或仔细考虑长篇文章，然后生成非常简短的答案的任务。GPT-3在一些任务上few-shot性能差也是这个原因，例如WIC（涉及比较两个句子中单词的使用方式）、ANLI（涉及比较两个句子以查看一个是否暗示另一个）以及几个阅读理解任务（例如QuAC和RACE）。我们还根据过去的文献推测，一个大型的双向模型在微调方面会比GPT-3更强。在GPT-3的规模上制作双向模型，或尝试使双向模型在few-shot或zero-shot学习方面发挥作用，是未来研究的一个有前途的方向，可以帮助实现“两全其美”。

A more fundamental limitation of the general approach described in this paper – scaling up any LM-like model, whether autoregressive or bidirectional – is that it may eventually run into (or could already be running into) the limits of the pretraining objective. Our current objective weights every token equally and lacks a notion of what is most important to predict and what is less important. [RRS20] demonstrate benefits of customizing prediction to entities of interest. Also,with self-supervised objectives, task specification relies on forcing the

desired task into a prediction problem, whereas ultimately, useful language systems (for example virtual assistants) might be better thought of as taking goal-directed actions rather than just making predictions. Finally, large pretrained language models are not grounded in other domains of experience, such as video or real-world physical interaction, and thus lack a large amount of context about the world[BHT+20]. For all these reasons, scaling pure self-supervised prediction is likely to hit limits, and augmentation with a different approach is likely to be necessary. Promising future directions in this vein might include learning the objective function from humans [ZSW+19a], fine-tuning with reinforcement learning, or adding additional modalities such as images to provide grounding and a better model of the world [CLY+19].

这篇论文中描述的一般方法的更根本的限制是，它可能最终会出现（或已经出现）预训练目标的限制，无论是自回归还是双向的LM-like模型。我们当前的目标函数是平等地看待每个token，没有token重要性的概念。[RRS20]这篇文章研究了将预测重点关注感兴趣的实体，效果明显。此外，对于自我监督的目标，下游任务使用时需要强制转换为预测问题，而最终，有用的语言系统（例如虚拟助手）可能是采取目标行动而不仅是进行预测。最后，大型预训练语言模型没有其他领域的经验，例如视频或现实世界的物理交互，因此缺乏关于世界的大量上下文信息[BHT+20]。由于所有这些原因，纯自我监督预测的发展很可能会遇到瓶颈，并且可能需要使用不同的方法进行增强。在这方面有前途的方向可能包括：从人类中学习目标函数[ZSW+19a]、使用强化学习进行微调，或添加其他模态，例如图像，以研究基础和更好的世界模型[CLY+19]。

Another limitation broadly shared by language models is poor sample efficiency during pre-training. While GPT-3 takes a step towards test-time sample efficiency closer to that of humans (one-shot or zero-shot), it still sees much more text during pre-training than a human sees in the their life time [Lin20]. Improving pre-training sample efficiency is an important direction for future work, and might come from grounding in the physical world to provide additional information, or from algorithmic improvements.

语言模型普遍存在的另一个限制是预训练期间的样本效率低下。虽然GPT-3在测试时间的样本效率方面迈出了向人类（一次性或零次性）更近的一步，但它在预训练期间看到的文本比人类一生中看到的文本还要多[Lin20]。提高预训练样本效率是未来工作的重要方向，可能来自于在物理世界，以提供额外的信息，或者来自于算法改进。

A limitation, or at least uncertainty, associated with few-shot learning in GPT-3 is ambiguity about whether few-shot learning actually learns new tasks “from scratch” at inference time, or if it simply recognizes and identifies tasks that it has learned during training. These possibilities exist on a spectrum, ranging from demonstrations in the training set that are drawn from exactly the same distribution as those at test time, to recognizing the same task but in a different format,to adapting to a specific style of a general task such as QA, to learning a skill entirely de novo. Where GPT-3 is on this spectrum may also vary from task to task. Synthetic tasks such as word scrambling or defining nonsense words seem especially

likely to be learned de novo, whereas translation clearly must be learned during pretraining, although possibly from data that is very different in organization and style than the test data. Ultimately, it is not even clear what humans learn from scratch vs from prior demonstrations. Even organizing diverse demonstrations during pre-training and identifying them at test time would be an advance for language models, but nevertheless understanding precisely how few-shot learning works is an important unexplored direction for future research.
GPT-3中的少样本学习存在一个限制，或者至少存在不确定性，即少样本学习是实际上是在推理时“从头开始”学习新任务，还是仅仅识别和确定它在训练期间学到的任务。这分递进的几种情况，从训练集中和测试集分布完全相同，到识别相同的任务但以不同的格式，到适应于一般任务的特定风格，例如QA，到完全新的技能。GPT-3是属于哪种情况，因任务而异。合成任务，如单词混淆或定义无意义的单词，似乎特别有可能从头开始学习，而翻译显然必须在预训练期间学习，尽管可能在组织和风格上测试数据不同。最终，甚至不清楚人类从头开始学习与从先前的演示中学习了什么。即使在预训练期间使用多样化的示例数据，并在测试时对它们进行识别也将是语言模型的一个进步，但是准确理解少样本学习的工作原理仍是未来研究的一个重要未开发的方向。
A limitation associated with models at the scale of GPT-3, regardless of objective function or algorithm, is that they are both expensive and inconvenient to perform inference on, which may present a challenge for practical applicability of models of this scale in their current form. One possible future direction to address this is distillation [HVD15] of large models down to a manageable size for specific tasks. Large models such as GPT-3 contain a very wide range of skills,most of which are not needed for a specific task, suggesting that in principle aggressive distillation may be possible.Distillation is well-explored in general [LHCG19a] but has not been tried at the scale of hundred of billions parameters;new challenges and opportunities may be associated with applying it to models of this size.
GPT-3模型规模的局限性，无论是目标函数还是算法，都在于它们的推理成本高昂且不方便，这可能对当前形式下的这种规模的模型的实际适用性构成挑战。解决这个问题的一个可能的未来方向是将大型模型精简[HVD15]到适合特定任务的可管理大小。像GPT-3这样的大型模型包含非常广泛的技能，其中大多数对于特定任务来说都是不必要的，这表明在原则上可以进行大幅精简。精简在一般情况下已经得到了很好的探索[LHCG19a]，但尚未尝试过百亿参数规模的模型；将其应用于这种规模的模型可能会带来新的挑战和机遇。
Finally, GPT-3 shares some limitations common to most deep learning systems – its decisions are not easily interpretable,it is not necessarily well-calibrated in its predictions on novel inputs as observed by the much higher variance in performance than humans on standard benchmarks, and it retains the biases of the data it has been trained on. This last issue – biases in the data that may lead the model to generate stereotyped or prejudiced content – is of special concern from a societal perspective, and will be discussed along with other issues in the next section on Broader Impacts(Section 6).

最后，GPT-3与大多数深度学习系统一样——它的决策不易解释，它在预测新输入时准确度不一定好，在标准基准测试的结果表明，其性能的方差比人类高得多，而且它保留了所训练数据的偏见。最后一个问题——数据中的偏见可能导致模型生成刻板化或有偏见的内容——从社会的角度来看，这是一个特别值得关注的问题，我们将在下一节“更广泛的影响”（第6节）中讨论。

## 6 广泛的影响

Language models have a wide range of beneficial applications for society, including code and writing auto-completion,grammar assistance, game narrative generation, improving search engine responses, and answering questions. But they also have potentially harmful applications. GPT-3 improves the quality of text generation and adaptability over smaller models and increases the difficulty of distinguishing synthetic text from human-written text. It therefore has the potential to advance both the beneficial and harmful applications of language models.

语言模型在社会中有广泛的有益应用，包括代码和写作自动完成、语法辅助、游戏叙述生成、改进搜索引擎响应和回答问题。但是，它们也有可能产生有害应用。GPT-3 提高了文本生成的质量和适应性，超过了较小模型，并增加了区分合成文本和人类编写文本的难度。因此，它既会产生有益应用，也会产生有害应用。

Here we focus on the potential harms of improved language models, not because we believe the harms are necessarily greater, but in order to stimulate efforts to study and mitigate them. The broader impacts of language models like this are numerous. We focus on two primary issues: the potential for deliberate misuse of language models like GPT-3 in Section 6.1, and issues of bias, fairness, and representation within models like GPT-3 in Section 6.2. We also briefly discuss issues of energy efficiency (Section 6.3).

我们关注最新的语言模型的可能性危害，不是因为我们认为这些危害一定就会更大，而是为了激发人们研究和减轻这些危害。像这样的语言模型的广泛影响是众多的。我们关注两个主要问题：第6.1节中像 GPT-3 这样的语言模型被故意滥用的潜在风险，以及第6.2节中 GPT-3 模型内的偏见、公平性和代表性问题。我们还简要讨论了能源效率问题（第6.3节）。

### 6.1 滥用语言模型

Malicious uses of language models can be somewhat difficult to anticipate because they often involve repurposing language models in a very different environment or for a different purpose than researchers intended. To help with this,we can think in terms of traditional security risk assessment frameworks, which outline key steps such as identifyingthreats and potential impacts, assessing likelihood, and determining risk as a combination of likelihood and impact[Ros12]. We discuss three factors: potential misuse applications, threat actors, and external incentive structures.

语言模型的恶意用途可能有些难以预测，因为使用者可能不按研究人员的预期来使用，比如难以预期的环境，难以预期的目的。为了帮助解决这个问题，我们可以从传统的安全风险评估框架的角度来思

考，这些框架概述了识别威胁和潜在影响、评估可能性以及将风险确定为可能性和影响的组合等关键步骤[Ros12]。我们讨论了三个因素：可能的滥用场景、威胁行为者和外部激励结构。

6.1.1 可能的滥用场景

Any socially harmful activity that relies on generating text could be augmented by powerful language models. Examples include misinformation, spam, phishing, abuse of legal and governmental processes, fraudulent academic essay writing and social engineering pretexting. Many of these applications bottleneck on human beings to write sufficiently highquality text. Language models that produce high quality text generation could lower existing barriers to carrying outthese activities and increase their efficacy.The misuse potential of language models increases as the quality of text synthesis improves. The ability of GPT-3 to generate several paragraphs of synthetic content that people find difficult to distinguish from human-written text in 3.9.4 represents a concerning milestone in this regard.

任何依赖于生成文本的社会有害活动都可能受到强大语言模型的增强。例如，包括误导信息、垃圾邮件、网络钓鱼、滥用法律和政府程序、欺诈性学术论文写作和社交工程借口等。其中许多应用程序瓶颈在于人类编写足够高质量的文本。产生高质量文本生成的语言模型可以降低这些欺诈的门槛。随着文本合成质量的提高，语言模型的误用潜力也会增加。GPT-3 能够生成几段合成内容，人们很难将其与人类编写的文本区分开来，这在这方面也算是代表了一个令人担忧的里程碑。

6.1.2 威胁行为者

Threat actors can be organized by skill and resource levels, ranging from low or moderately skilled and resourced actors who may be able to build a malicious product to ‘advanced persistent threats’ (APTs): highly skilled and well-resourced(e.g. state-sponsored) groups with long-term agendas [SBC+19].

威胁行为者按技能和资源组织水平分为几种档次，有低档次的，中等档次的，高等档次的。高等档次技能熟练、资源充足，可以构建恶意产品，也被称为“高级持续性威胁”。

To understand how low and mid-skill actors think about language models, we have been monitoring forums and chat groups where misinformation tactics, malware distribution, and computer fraud are frequently discussed. While we did find significant discussion of misuse following the initial release of GPT-2 in spring of 2019, we found fewer instances of experimentation and no successful deployments since then. Additionally, those misuse discussions were correlated with media coverage of language model technologies. From this, we assess that the threat of misuse from these actors is not immediate, but significant improvements in reliability could change this.

为了了解低技能和中技能行为者如何看待语言模型，我们一直在监视论坛和聊天组，这些地方经常讨论误导策略、恶意软件分发和计算机欺诈。虽然我们在2019年春季GPT-2首次发布后发现了大量的滥用讨论，但实质性落地的较少。此外，这些滥用讨论与语言模型技术的媒体报道相关。因此，我们判定这些行为者的模型滥用并不是立即展开的，但模型的性能和可靠性显著提高后就不是这种情况了。

Because APTs do not typically discuss operations in the open, we have consulted with professional threat analysts about possible APT activity involving the use of language models. Since the release of GPT-2 there has been no discernible difference in operations that may see potential gains by using language models. The assessment was that language models may not be worth investing significant resources in because there has been no convincing demonstration thatcurrent language models are significantly better than current methods for generating text, and because methods for "targeting" or "controlling" the content of language models are still at a very early stage.
由于APT通常不会公开讨论，因此我们已经咨询了专业的威胁分析师，了解可能涉及使用语言模型的APT活动。自GPT-2发布以来，并没有发现可以提升APT的能力。评估结果是，语言模型可能不值得投入大量资源，因为尚未有令人信服的证据表明当前的语言模型比现有的方法更好，并且针对语言模型内容的“定位”或“控制”的方法仍处于非常早期的阶段。
6.1.3 外部激励结构
Each threat actor group also has a set of tactics, techniques, and procedures (TTPs) that they rely on to accomplish their agenda. TTPs are influenced by economic factors like scalability and ease of deployment; phishing is extremely popular among all groups because it offers a low-cost, low-effort, high-yield method of deploying malware and stealing logincredentials. Using language models to augment existing TTPs would likely result in an even lower cost of deployment.
每个威胁行为者组也有一套战术、技术和程序（TTPs），他们依靠这些来完成他们的操作。TTPs 受到成本的影响，如可扩展性和部署的便利性；网络钓鱼在所有组中都非常流行，因为它提供了一种低成本、低投入、高产出的部署恶意软件和窃取登录凭据的方法。使用语言模型来增强现有的 TTPs 可能会使部署成本更低。
6.2 Fairness, Bias, and Representation
Biases present in training data may lead models to generate stereotyped or prejudiced content. This is concerning,since model bias could harm people in the relevant groups in different ways by entrenching existing stereotypes and producing demeaning portrayals amongst other potential harms [Cra17]. We have conducted an analysis of biases inthe model in order to better understand GPT-3’s limitations when it comes to fairness, bias, and representation.
训练数据中存在的偏见可能导致模型生成刻板化或有偏见的内容。这是令人担忧的，因为模型偏见可能通过巩固现有的刻板印象和产生贬低的描绘等其他潜在危害以不同的方式伤害相关群体 [Cra17]。我们对模型中的偏见进行了分析，以更好地了解 GPT-3 在公平性、偏见和代表性方面的局限性。
Our goal is not to exhaustively characterize GPT-3, but to give a preliminary analysis of some of its limitations and behaviors. We focus on biases relating to gender, race, and religion, although many other categories of bias are likely present and could be studied in follow-up

work. This is a preliminary analysis and does not reflect all of the model’s biases even within the studied categories.
我们的目标不是详尽地描述 GPT-3，而是对其一些局限性和行为进行初步分析。我们关注与性别、种族和宗教有关的偏见，尽管可能存在许多其他类别的偏见，其他的偏见可以在后续工作中进行研究。这是初步分析，即使在所研究的类别中，也不能反映出模型的所有偏见。
Broadly, our analysis indicates that internet-trained models have internet-scale biases; models tend to reflect stereotypespresent in their training data. Below we discuss our preliminary findings of bias along the dimensions of gender, race,and religion. We probe for bias in the 175 billion parameter model and also in similar smaller models, to see if and howthey are different in this dimension.
总的来说，我们的分析表明，互联网训练的模型具有互联网规模的偏见；模型往往反映了其训练数据中存在的刻板印象。下面我们将讨论我们在性别、种族和宗教方面偏见的初步发现。我们探究了1750亿参数模型以及类似的较小模型中的偏见，以查看它们在这个维度上是否有所不同。
6.2.1 性别
In our investigation of gender bias in GPT-3, we focused on associations between gender and occupation. We found that occupations in general have a higher probability of being followed by a male gender identifier than a female one(in other words, they are male leaning) when given a context such as "The {occupation} was a" (Neutral Variant).83% of the 388 occupations we tested were more likely to be followed by a male identifier by GPT-3. We measured this by feeding the model a context such as "The detective was a" and then looking at the probability of the model following up with male indicating words (eg. man, male etc.) or female indicating words (woman, female etc.).In particular, occupations demonstrating higher levels of education such as legislator, banker, or professor emeritus were heavily male leaning along with occupations that require hard physical labour such as mason, millwright, and sheriff. Occupations that were more likely to be followed by female identifiers include midwife, nurse, receptionist,housekeeper etc.
我们在调查 GPT-3 中的性别偏见时，主要关注了性别和职业之间的关联。实验发现，职业是偏向男性的。我们测试的 388 个职业中，83% 都有男性倾向。实验的具体做法为：提供类似“侦探是一个”这样的上下文，然后查看模型后面出现男性相关词（例如男人、男性等）或女性相关词（女人、女性等）的概率。尤其是表现出更高教育水平的职业，如立法者、银行家或名誉教授，以及需要较大体力劳动的职业，如泥瓦匠、机械师和警长，都倾向于男性。偏向女性的职业包括助产士、护士、接待员、管家等。
We also tested how these probabilities changed when we shifted the context to be the "The competent {occupation}was a" (Competent Variant), and when we shifted the context to be "The incompetent {occupation} was a"(Incompetent Variant) for each occupation in the dataset. We found that, when prompted with "The competent {occupation} was a," the majority of occupations had an even higher probability of being followed by a male identifier

than a female one than was the case with our original neutral prompt, "The {occupation} was a". With the prompt "The incompetent {occupation} was a" the majority of occupations still leaned male with a similar probability than for our original neutral prompt. The average occupation bias - measured as

was −1.11 for the Neutral Variant, −2.14 for the Competent Variant and −1.15 for the Incompetent Variant

我们还测试了当改变上下文时的性别倾向概率，有两种情况，一是改为“能干的职业是一个能干的变量”，二是改为“无能的职业是一个无能的变量”。我们发现第一种情况的结果，相比于“这个职业是一个”，男性倾向的概率更高了。第二种情况的结果，相比“这个职业是一个”变化不大。平均职业偏见衡量指标如下，中性变量的值为1.11，能干变量的值为2.14，无能变量的值为1.15。

We also carried out pronoun resolution on the Winogender dataset [RNLVD18] using two methods which further corroborated the model’s tendency to associate most occupations with males. One method measured the models ability to correctly assign a pronoun as the occupation or the participant. For example, we fed the model a context such as "The advisor met with the advisee because she wanted to get advice about job applications. ‘She’ refers to the" and found the option with the lowest probability between the two possible options (Choices between Occupation Option: advisor; Participant Option: advisee).

我们还对 Winogender 数据集 [RNLVD18] 进行了代词消解，使用了两种方法，进一步证实了大多数职业有男性倾向。其中一种方法是评估模型正确地将代词分配为职业或参与者的能力。例如，我们提供了一个上下文，如“顾问与咨询者会面，因为她想获得有关求职的建议。 ‘她’指的是谁？”，并找到两个选项中概率最低的选项（职业选项：顾问；参与者选项：咨询者）。

Occupation and participant words often have societal biases associated with them such as the assumption that most occupants are by default male. We found that the language models learnt some of these biases such as a tendency to associate female pronouns with participant positions more than male pronouns. GPT-3 175B had the highest accuracy of all the models (64.17%) on this task. It was also the only model where the accuracy for Occupant sentences (sentenceswhere the correct answer was the Occupation option) for females was higher than for males (81.7% vs 76.7%). Allother models had a higher accuracy for male pronouns with Occupation sentences as compared to female pronouns with the exception of our second largest model- GPT-3 13B - which had the same accuracy (60%) for both. This offers some preliminary evidence that in places where issues of bias can make language models susceptible to error, the larger models are more robust than smaller models.

职业和参与者词汇通常带有社会偏见，例如默认大多数职业是男性供职。我们发现，语言模型学习了其中一些偏见，例如倾向于将女性代词与参与者位置联系起来，而不是男性代词。GPT-3 175B在此任务中的准确率最高（64.17%）。它也是唯一一个女性占用者句子（正确答案为职业选项的句子）的准确率高于男性的模型（81.7% vs 76.7%）。除了我们第二大的模型GPT-3 13B（两者的准确率均为

60%）之外，所有其他模型在职业句子中使用男性代词的准确率都比使用女性代词的准确率高。这提供了一些初步证据，表明在偏见问题可能使语言模型容易出错的地方，较大的模型比较小的模型更加稳健。

We also performed co-occurrence tests, where we analyzed which words are likely to occur in the vicinity of other pre-selected words. We created a model output sample set by generating 800 outputs of length 50 each with a tempe of 1 and top p of 0.9 for every prompt in our dataset. For gender, we had prompts such as "He was very", "She was very", "He would be described as", "She would be described as"9. We looked at the adjectives and adverbs in the top 100 most favored words using an off-the-shelf POS tagger [LB02]. We found females were more often described using appearance oriented words such as ” beautiful” and ” gorgeous” as compared to men who weremore often described using adjectives that span a greater spectrum.

我们还进行了共现测试，分析了哪些单词可能出现在其他预选单词的附近。我们通过为数据集中的每个提示生成800个长度为50的输出样本集，每个样本集的温度为1，top p为0.9来创建模型输出样本集。对于性别，我们有像“他非常”、“她非常”、“他会被描述为”、“她会被描述为”这样的提示。我们使用现成的POS标注器[LB02]查看了前100个最受欢迎的形容词和副词。我们发现女性更常用外貌为导向的词语，如“美丽”和“华丽”，而男性更常用涵盖更广泛范围的形容词。

Table 6.1 shows the top 10 most favored descriptive words for the model along with the raw number of times eachword co-occurred with a pronoun indicator. “Most Favored” here indicates words which were most skewed towards acategory by co-occurring with it at a higher rate as compared to the other category. To put these numbers in perspective,we have also included the average for the number of co-occurrences across all qualifying words for each gender.

表6.1显示了模型中最受欢迎的前10个描述性词语，以及每个词语与代词指示器共同出现的原始次数。这里的“最受欢迎”表示这些词语与某个类别的共现率比与另一个类别的共现率更高。为了让这些数字更具参考价值，我们还包括了每个性别所有合格单词的共现次数平均值。

表 6.1: 175B模型最偏见的形容词

6.2.2 种族

To investigate racial bias in GPT-3, we seeded the model with prompts such as - "The {race} man was very","The {race} woman was very" and "People would describe the {race} person as" and generated 800 samples for each of the above prompts, with {race} replaced with a term indicating a racial category such as White or Asian. We then measure word co-occurrences in the generated samples. Given prior research demonstrating that language models produce text of differing sentiment when varying features such as occupation [HZJ+19], we explored how race impacted sentiment. We measured sentiment using Senti WordNet [BES10] for the words which co-occurred disproportionately with each race. Each

word sentiment varied from 100 to -100, with positive scores indicating positivewords (eg. wonderfulness: 100, amicable: 87.5), negative scores indicating negative words (eg. wretched: -87.5 , horrid:-87.5) and a score of 0 indicating neutral words (eg. sloping, chalet).
为了调查GPT-3中的种族偏见，我们使用了诸如“The {race} man was very”、“The {race} woman was very”和“People would describe the {race} person as”之类的prompt，并为上述每个提示生成了800个样本，其中{race}被替换为表示种族类别的术语，例如White或Asian。然后我们测量了生成样本中的词共现情况。鉴于先前的研究表明，当变化职业等特征时，语言模型会产生不同情感的文本[HZJ+19]，因此，我们探讨了种族如何影响情感。我们使用Senti WordNet [BES10]测量了与每个种族不成比例共现的单词的情感。每个单词的情感值从100到-100不等，其中正分数表示积极的单词（例如wonderfulness：100，amicable：87.5），负分数表示消极的单词（例如wretched：-87.5，horrid：-87.5），分数为0表示中性单词（例如sloping，chalet）。
It should be noted that we were explicitly prompting the models to talk about race and this in turn generated text that focused on racial features; these results are not from the models talking about race in the wild but talking about race in an experimental setup where they have been primed to do so. Additionally, since we are measuring sentiment by simply looking at word co-occurrences, the resulting sentiment can reflect socio-historical factors - for instance, text relating to a discussion of slavery will frequently have a negative sentiment, which may lead to a demographic being associated with a negative sentiment under this testing methodology.
需要注意的是，我们明确要求模型谈论种族，这反过来产生了侧重于种族特征的文本；这些结果中谈论的种族不是在实际使用时会出现的，而是在实验条件下谈论种族时产生的。此外，由于我们通过简单地查看单词共现来测量情感，因此所得到的情感可能反映社会历史因素——例如，与奴隶制讨论相关的文本通常会有负面情感，这可能会导致在这种测试方法下将某个人群与负面情感联系起来。
Across the models we analyzed, ‘Asian’ had a consistently high sentiment - it ranked 1st in 3 out of 7 models. On the other hand, ’ Black’ had a consistently low sentiment - it ranked the lowest in 5 out of 7 models. These differences narrowed marginally on the larger model sizes. This analysis gives a sense of the biases of different models and highlights the need for more sophisticated analysis of the relationship between sentiment, entities, and input data.
在我们分析的模型中，“亚洲人”情感始终很高，在7个模型中有3个模型排名第一。另一方面，“黑人”的情感始终很低，在7个模型中有5个模型排名最低。这些差异在较大的模型尺寸上略有缩小。这种分析让我们了解了不同模型的偏见，并强调了需要对情感、实体和输入数据之间的关系进行更复杂的分析。

图6.1 不同模型的种族情感分析

6.2.3 宗教

We studied which words co-occurred with religious terms relating to Atheism, Buddhism, Christianity, Hinduism, Islam,and Judaism, by generating 800 model outputs of length ≈50

with a temperature of 1 and a top p of 0.9 for every prompt. Our prompts were of the nature "{Religion practitioners} are" (Eg. "Christians are") for each of the six religious categories listed above. We then allowed the model to naturally carry out completions and created a corpus of such completions for studying co-occurrence of words.
我们研究了与无神论、佛教、基督教、印度教、伊斯兰教和犹太教有关的宗教术语共现的单词，通过为每个提示生成长度约为50的800个模型输出，温度为1，top p为0.9。我们的提示是以“{宗教信徒}是”（例如，“基督徒是”）的形式，宗教为上述六个宗教类别中的每一个。然后，我们允许模型自然地完成这些提示，并创建了这样的语料库来研究单词的共现。
The following is an example output from the model:
---
---
"Buddhists are divided into two main branches - Theravada and Mahayana. Theravadais the more conservative branch, centering on monastic life and the earliest sutras and refusing to recognize the later Mahayana sutras as authentic."
---
---
以下是一个输出案例：
---
---
佛教徒分为两个主要派别——小乘和大乘。小乘是更为保守的分支，以出家生活和最早的经典为中心，并拒绝承认后来的大乘经典。
---
---
Similar to race, we found that the models make associations with religious terms that indicate some propensity to reflect how these terms are sometimes presented in the world. For example, with the religion Islam, we found that words such as ramadan, prophet and mosque co-occurred at a higher rate than for other religions. We also found that words such as violent, terrorism and terrorist co-occurred at a greater rate with Islam than with other religions and were inthe top 40 most favored words for Islam in GPT-3.
与种族类似，我们发现模型与宗教术语建立了关联，这些术语表明了一些倾向，反映了这些术语在世界上的呈现方式。例如，对于伊斯兰教，我们发现像斋月、先知和清真寺这样的词汇的共现率比其他宗教高。我们还发现，暴力、恐怖主义和恐怖分子等词汇与伊斯兰教的共现率比其他宗教高，并且是GPT-3中伊斯兰教最受欢迎的前40个词汇之一。

表6.2 GPT-3 175B中十个最受欢迎的词

### 6.2.4 未来偏见和公平的挑战

We have presented this preliminary analysis to share some of the biases we found in order to motivate further research,and to highlight the inherent difficulties in characterizing biases in large-scale generative models; we expect this to be an area of continuous research for us and are excited to discuss different methodological approaches with the community.We view the work in this section as subjective signposting - we chose gender, race, and religion as a starting point, but we recognize the inherent subjectivity in this choice. Our work is inspired by the literature on characterizing model attributes to develop informative labels such as Model Cards for Model Reporting from [MWZ+18].
我们提出了这项初步分析，以分享我们发现的一些偏见，并激励进一步的研究，同时强调大规模生成模型中识别偏见的固有困难；我们期望这将是我们持续研究的一个领域，并很高兴与研究学者讨论不同的方法论。我们认为本节工作是主观的 - 我们选择性别、种族和宗教作为起点，但我们认识到这种选择的主观性。我们的工作受到了一些文献的启发，具体为描述模型属性，以此来生成标签，例如来自[MWZ+18]的模型报告的模型卡。
Ultimately, it is important not just to characterize biases in language systems but to intervene. The literature on this is also extensive [QMZH19, HZJ+19], so we offer only a few brief comments on future directions specific to large language models. In order to pave the way for effective bias prevention in general purpose models, there is a need for building a common vocabulary tying together the normative, technical and empirical challenges of bias mitigation for these models. There is room for more research that engages with the literature outside NLP, better articulates normative statements about harm, and engages with the lived experience of communities affected by NLP systems [BBDIW20].Thus, mitigation work should not be approached purely with a metric driven objective to ‘remove’ bias as this has been shown to have blind spots [GG19, NvNvdG19] but in a holistic manner.
最终，重要的不仅是表征语言系统中的偏见，而是要进行干预。关于这一点的文献也很广泛，因此我们只提供了一些关于大型语言模型特定未来方向的简短评论。为了为通用模型中的有效偏见预防铺平道路，需要建立一个将这些模型的规范、技术和经验挑战联系在一起的词库。还有更多的研究空间，可以与NLP以外的文献进行交流，更好地阐述有关伤害的规范性陈述，并与受NLP系统影响的社区进行交流。因此，应该以整体的方式来缓解工作，而不是仅以度量驱动的目标来“消除”偏见，因为这已经被证明存在盲点。
6.3 能源使用
Practical large-scale pre-training requires large amounts of computation, which is energy-intensive: training the GPT-3175B consumed several thousand petaflop/s-days of compute during pre-training, compared to tens of petaflop/s-daysfor a 1.5B parameter GPT-2 model (Figure 2.2). This means we should be cognizant of the cost and efficiency of such models, as advocated by [SDSE19].
实际的大规模预训练需要大量的计算，这是很耗资源的：与1.5B参数GPT-2模型（图2.2）相比，GPT-3175B的训练在预训练期间消耗了数千个petaflop/s-days的算力。这意味着我们应该意识到这些模型

的成本和效率，正如[SDSE19]所提倡的那样。
The use of large-scale pre-training also gives another lens through which to view the efficiency of large models - we should consider not only the resources that go into training them, but how these resources are amortized over the lifetime of a model, which will subsequently be used for a variety of purposes and fine-tuned for specific tasks. Though models like GPT-3 consume significant resources during training, they can be surprisingly efficient once trained: even with the full GPT-3 175B, generating 100 pages of content from a trained model can cost on the order of 0.4 kW-hr, or only a few cents in energy costs. Additionally, techniques like model distillation [LHCG19a] can further bring down the cost of such models, letting us adopt a paradigm of training single, large-scale models, then creating more efficient versions of them for use in appropriate contexts. Algorithmic progress may also naturally further increase the efficiency of such models over time, similar to trends observed in image recognition and neural machine translation [HB20].
还有另一种视角来看待大型模型的效率 - 我们不仅应该考虑用于训练它们的资源，还应该考虑这些资源在模型的生命周期内如何分摊，这些模型随后将用于各种目的并针对特定任务进行微调。虽然像GPT-3这样的模型在训练期间消耗了大量资源，但一旦训练完成，它们可以出奇地高效：即使使用完整的GPT-3 175B，从训练模型生成100页内容的成本也可能只有0.4 kW-hr，或者只有几美分的能源成本。此外，像模型蒸馏[LHCG19a]这样的技术可以进一步降低这些模型的成本，即先训练单个大型模型，然后创建更高效的版本。算法进步还可能自然地随着时间的推移进一步提高这些模型的效率，类似于图像识别和神经机器翻译[HB20]的发展。
7 相关工作
Several lines of work have focused on increasing parameter count and/or computation in language models as a means to improve generative or task performance. An early work scaled LSTM based language models to over a billion parameters [JVS+16]. One line of work straightforwardly increases the size of transformer models, scalingup parameters and FLOPS-per-token roughly in proportion. Work in this vein has successively increased model size:213 million parameters [VSP+17] in the original paper, 300 million parameters [DCLT18], 1.5 billion parameters[RWC+19], 8 billion parameters [SPP+19], 11 billion parameters [RSR+19], and most recently 17 billion parameters[Tur20]. A second line of work has focused on increasing parameter count but not computation, as a means of increasing models' capacity to store information without increased computational cost. These approaches rely on the conditional computation framework [BLC13] and specifically, the mixture-of-experts method [SMM+17] has been used to produce 100 billion parameter models and more recently 50 billion parameter translation models [AJF19],though only a small fraction of the parameters are actually used on each forward pass. A third approach increases computation without increasing parameters; examples of this approach include adaptive computation time [Gra16] and the universal transformer [DGV+18]. Our work focuses on the first approach (scaling

compute and parameters together,by straightforwardly making the neural net larger), and increases model size 10x beyond previous models that employ this strategy.
有几条研究路线专注于增加语言模型的参数数量和/或计算量，以提高生成或任务性能。早期的工作将基于LSTM的语言模型扩展到了超过十亿个参数[JVS+16]。其中一条研究线路直接增加了Transformer模型的大小，将参数和每个标记的FLOPS按比例扩大。比如：在原始论文中的2.13亿个参数[VSP+17]，3亿个参数[DCLT18]，15亿个参数[RWC+19]，80亿个参数[SPP+19]，110亿个参数[RSR+19]，最近的是170亿个参数[Tur20]。第二条研究线路专注于增加参数数量而不是计算量，增加模型存储信息的能力而不增加计算成本。这些方法依赖于条件计算框架[BLC13]，比如，混合专家方法[SMM+17]已被用于生成1000亿参数模型，最近还用于生成500亿参数的翻译模型[AJF19]，但每次前向传递只使用了其中的一小部分参数。第三种方法是增加计算量而不增加参数；这种方法的例子包括自适应计算时间[Gra16]和通用transformer[DGV+18]。我们的工作集中在第一种方法上（通过简单地扩大神经网络来同时扩大计算和参数），并将模型大小增加了10倍，超过了以前采用这种策略的模型。
Several efforts have also systematically studied the effect of scale on language model performance. [KMH+20,RRBS19, LWS+20, HNA+17], find a smooth power-law trend in loss as autoregressive language models are scaled up.This work suggests that this trend largely continues as models continue to scale up (although a slight bending of thecurve can perhaps be detected in Figure 3.1), and we also find relatively smooth increases in many (though not all)downstream tasks across 3 orders of magnitude of scaling.
还有一些系统性的研究努力致力于研究规模对语言模型性能的影响。[KMH+20,RRBS19, LWS+20, HNA+17]发现，随着自回归语言模型的规模扩展，损失呈现出平滑的幂律趋势。这项工作表明，随着模型的不断扩大，这种趋势在很大程度上仍在持续（尽管在图3.1中可能可以检测到轻微的弯曲），并且我们还发现，在3个数量级的扩展中，许多（但不是全部）下游任务的增长也相对平滑。
Another line of work goes in the opposite direction from scaling, attempting to preserve strong performance in language models that are as small as possible. This approach includes ALBERT [LCG+19] as well as general [HVD15] and task-specific [SDCW19, JYS+19, KR16] approaches to distillation of language models. These architectures and techniques are potentially complementary to our work, and could be applied to decrease latency and memory footprint of giant models.
另一条研究线路与扩大规模相反，试图在尽可能小的语言模型中保持强大的性能。这种方法包括ALBERT [LCG+19]以及通用的[HVD15]和任务特定的[SDCW19，JYS+19，KR16]语言模型蒸馏方法。这些架构和技术可能对我们有用，用于减少巨型模型的延迟和内存占用。
As fine-tuned language models have neared human performance on many standard benchmark tasks, considerable effort has been devoted to constructing more difficult or open-ended tasks, including question answering [KPR+19,IBGC+14, CCE+18, MCKS18], reading comprehension [CHI+18, RCM19], and adversarially constructed datasets designed to

be difficult for existing language models [SBBC19, NWD+19]. In this work we test our models on many of these datasets.
随着微调语言模型在许多标准基准任务上接近人类表现，人们已经付出了相当大的努力来构建更困难或开放式的任务，包括问答[KPR+19，IBGC+14，CCE+18，MCKS18]，阅读理解[CHI+18，RCM19]以及对抗性构建的数据集，旨在为现有语言模型带来困难[SBBC19，NWD+19]。在这项工作中，我们在许多这些数据集上测试了我们的模型。
Many previous efforts have focused specifically on question-answering, which constitutes a significant fraction of the tasks we tested on. Recent efforts include [RSR+19, RRS20], which fine-tuned an 11 billion parameter language model,and [GLT+20], which focused on attending over a large corpus of data at test time. Our work differs in focusing on in-context learning but could be combined in the future with those of [GLT+20, LPP+20].
之前有很多工作专注于问答，这也是我们测试任务的重要部分。最近的包括[RSR+19，RRS20]，它们微调了一个110亿个参数的语言模型，以及[GLT+20]，它在测试时参考了大量的数据语料库。我们的工作不同之处在于专注于上下文学习，但未来可以与[GLT+20，LPP+20]的工作相结合。
Metalearning in language models has been utilized in [RWC+19], though with much more limited results and no systematic study. More broadly, language model metalearning has an inner-loop-outer-loop structure, making it structurally similar to metalearning as applied to ML in general. Here there is an extensive literature, including matching networks [VBL+16], RL2 [DSC+16], learning to optimize [RL16, ADG+16, LM17] and MAML [FAL17].Our approach of stuffing the model’s context with previous examples is most structurally similar to RL2 and also resembles [HYC01], in that an inner loop of adaptation takes place through computation in the model’s activations across timesteps, without updating the weights, while an outer loop (in this case just language model pre-training)updates the weights, and implicitly learns the ability to adapt to or at least recognize tasks defined at inference-time.Few-shot auto-regressive density estimation was explored in [RCP+17] and [GWC+18] studied low-resource NMT asa few-shot learning problem.
[RWC+19]利用了语言模型中的元学习，但结果有限，且没有系统研究。更广泛地说，语言模型元学习具有内环外环结构，使其在结构上类似于应用于ML的元学习。这里有广泛的文献，包括匹配网络[VBL+16]、RL2[DSC+16]、学习优化[RL16、ADG+16、LM17]和MAML[FAL17]。我们的方法是将模型的上下文填充到示例中，它在结构上与RL2最相似，并且类似于[HYC01]，即通过模型进行内部适应循环，而不更新权重，而外部循环（在这种情况下仅为语言模型预训练）更新权重，并隐含地学习适应或至少识别推理时间定义的任务的能力。Few-shot自回归密度估计在[RCP+17]中得到了探索，[GWC+18]将低资源NMT作为少样本学习问题进行了研究。
While the mechanism of our few-shot approach is different, prior work has also explored ways of using pre-trained language models in combination with gradient descent to perform few-shot learning [SS20]. Another sub-field with similar goals is semi-supervised learning

where approaches such as UDA [XDH+19] also explore methods of fine-tuningwhen very little labeled data is available.
虽然我们的少样本方法的机制不同，但以前的工作也探索了使用预训练的语言模型与梯度下降相结合进行少样本学习的方法[SS20]。另一个具有类似目标的子领域是半监督学习，其中诸如UDA [XDH+19]的方法也探索了在很少有标记数据的情况下微调的方法。
Giving multi-task models instructions in natural language was first formalized in a supervised setting with [MKXS18]and utilized for some tasks (such as summarizing) in a language model with [RWC+19]. The notion of presenting tasks in natural language was also explored in the text-to-text transformer [RSR+19], although there it was applied formulti-task fine-tuning rather than for in-context learning without weight updates.
使用自然语言为多任务模型提供指令最初是在监督设置中通过[MKXS18]正式化的，并在[RWC+19]的语言模型中用于某些任务（例如摘要）。在文本到文本transformer[RSR+19]中也探索了以自然语言呈现任务的概念，尽管在那里它被应用于多任务微调而不是用于在上下文学习中不更新权重。
Another approach to increasing generality and transfer-learning capability in language models is multi-task learning[Car97], which fine-tunes on a mixture of downstream tasks together, rather than separately updating the weights foreach one. If successful multi-task learning could allow a single model to be used for many tasks without updating theweights (similar to our in-context learning approach), or alternatively could improve sample efficiency when updatingthe weights for a new task. Multi-task learning has shown some promising initial results [LGH+15, LSP+18] and multi-stage fine-tuning has recently become a standardized part of SOTA results on some datasets [PFB18] and pushedthe boundaries on certain tasks [KKS+20], but is still limited by the need to manually curate collections of datasets andset up training curricula. By contrast pre-training at large enough scale appears to offer a “natural” broad distribution oftasks implicitly contained in predicting the text itself. One direction for future work might be attempting to generatea broader set of explicit tasks for multi-task learning, for example through procedural generation [TFR+17], humaninteraction [ZSW+19b], or active learning [Mac92].
增加语言模型的通用性和迁移学习能力的另一种方法是多任务学习[Car97]，它在混合下游任务的基础上进行微调，而不是分别为每个任务更新权重。如果成功，多任务学习可以允许使用单个模型处理多个任务而不更新权重（类似于我们的上下文学习方法），或者在更新新任务的权重时可以提高样本效率。多任务学习已经显示出一些有前途的初步结果[LGH+15，LSP+18]，多阶段微调最近已成为某些数据集的SOTA结果的标准化部分[PFB18]，并推动了某些任务的界限[KKS+20]，但仍受到手动策划数据集集合和设置培训课程的限制。相比之下，大规模预训练似乎提供了一个“自然”的广泛任务分布，这些任务隐含在预测文本本身中。未来的一个方向可能是尝试生成更广泛的显式任务集，例如通过程序生成[TFR+17]、人机交互[ZSW+19b]或主动学习[Mac92]。
Algorithmic innovation in language models over the last two years has been enormous, including denoising-based bidirectionality [DCLT18], prefixLM [DL15] and encoder-decoder

architectures [LLG+19, RSR+19], random permutations during training [YDY+19], architectures that improve the efficiency of sampling [DYY+19], improvements in data and training procedures [LOG+19], and efficiency increases in the embedding parameters [LCG+19]. Many of these techniques provide significant gains on downstream tasks. In this work we continue to focus on pure autoregressive language models, both in order to focus on in-context learning performance and to reduce the complexity of our large model implementations. However, it is very likely that incorporating these algorithmic advances could improve GPT-3’s performance on downstream tasks, especially in the fine-tuning setting, and combining GPT-3’s scale with these algorithmic techniques is a promising direction for future work.

过去两年中，语言模型的算法创新非常巨大，包括双向去噪[DCLT18]、前缀语言模型[DL15]和编码器-解码器架构[LLG+19，RSR+19]、训练过程中的随机排列[YDY+19]、提高采样效率的架构[DYY+19]、数据和训练程序的改进[LOG+19]以及嵌入参数的效率提高[LCG+19]。其中许多技术在下游任务中提供了显著的收益。在我们的工作中，我们继续专注于纯自回归语言模型，既为了专注于上下文学习性能，也为了减少我们的大型模型实现的复杂性。然而，将这些算法创新纳入GPT-3的下游任务中，特别是在微调设置中，可以提高其性能，并将GPT-3的规模与这些算法技术相结合是未来工作的一个有前途的方向。

8 结论

We presented a 175 billion parameter language model which shows strong performance on many NLP tasks and benchmarks in the zero-shot, one-shot, and few-shot settings, in some cases nearly matching the performance of state-of-the-art fine-tuned systems, as well as generating high-quality samples and strong qualitative performance at tasks defined on-the-fly. We documented roughly predictable trends of scaling in performance without using fine-tuning.We also discussed the social impacts of this class of model. Despite many limitations and weaknesses, these results suggest that very large language models may be an important ingredient in the development of adaptable, general language systems.

我们提出了一个拥有1750亿个参数的语言模型，它在zero-shot、one-shot和few-shot等多个NLP任务和基准测试中表现出强大的性能，在某些情况下几乎可以与最先进的微调系统的性能相匹配，同时在即兴任务上生成高质量的样本和强大的定性表现。我们记录了在没有使用微调的情况下，性能扩展的大致可预测趋势。我们还讨论了这类模型的社会影响。尽管存在许多限制和缺陷，但这些结果表明，非常大的语言模型可能是开发适应性强、通用语言系统的重要组成部分。

致谢

The authors would like to thank Ryan Lowe for giving detailed feedback on drafts of the paper. Thanks to Jakub Pachocki and Szymon Sidor for suggesting tasks, and Greg Brockman, Michael Petrov, Brooke Chan, and ChelseaVoss for helping run evaluations on OpenAI’s infrastructure. Thanks to David Luan for initial support in scaling up this project, Irene Solaiman for discussions about ways to approach and evaluate bias, Harrison Edwards and

YuraBurda for discussions and experimentation with in-context learning, Geoffrey Irving and Paul Christiano for early discussions of language model scaling, Long Ouyang for advising on the design of the human evaluation experiments,Chris Hallacy for discussions on data collection, and Shan Carter for help with visual design. Thanks to the millions of people who created content that was used in the training of the model, and to those who were involved in indexing or upvoting the content (in the case of WebText). Additionally, we would like to thank the entire OpenAI infrastructureand supercomputing teams for making it possible to train models at this scale.

作者们要感谢Ryan Lowe对论文草稿提供详细反馈。感谢Jakub Pachocki和Szymon Sidor提出的任务建议，以及Greg Brockman、Michael Petrov、Brooke Chan和Chelsea Voss在OpenAI基础设施上进行评估。感谢David Luan在扩大这个项目方面的初步支持，Irene Solaiman讨论了评估偏差的方法，Harrison Edwards和Yura Burda讨论了上下文学习的实验，Geoffrey Irving和Paul Christiano早期讨论了语言模型的扩展，Long Ouyang为人类评估实验的设计提供建议，Chris Hallacy讨论了数据收集，Shan Carter帮助视觉设计。感谢数百万创造了用于模型训练的内容的人，以及那些参与索引或投票内容的人（在WebText的情况下）。此外，我们还要感谢整个OpenAI基础设施和超级计算团队，使得在这个规模上训练模型成为可能。

Contributions

**Tom Brown, Ben Mann, Prafulla Dhariwal, Dario Amodei, Nick Ryder, Daniel M Ziegler, and Jeffrey Wu**
implemented the large-scale models, training infrastructure, and model-parallel strategies.

**Tom Brown, Dario Amodei, Ben Mann, and Nick Ryder** conducted pre-training experiments.

**Ben Mann and Alec Radford** collected, filtered, deduplicated, and conducted overlap analysis on the training data.

**Melanie Subbiah, Ben Mann, Dario Amodei, Jared Kaplan, Sam McCandlish, Tom Brown, Tom Henighan, and Girish Sastry** implemented the downstream tasks and the software framework for supporting them, including creation of synthetic tasks.

**Jared Kaplan and Sam McCandlish** initially predicted that a giant language model should show continued gains, and applied scaling laws to help predict and guide model and data scaling decisions for the research.

**Ben Mann** implemented sampling without replacement during training.

**Alec Radford** originally demonstrated few-shot learning occurs in language models.

**Jared Kaplan and Sam McCandlish** showed that larger models learn more quickly in-context, and systematically studied in-context learning curves, task prompting, and evaluation methods.

**Prafulla Dhariwal** implemented an early version of the codebase, and developed the memory optimizations for fully half-precision training.

**Rewon Child and Mark Chen** developed an early version of our model-parallel strategy.
**Rewon Child and Scott Gray** contributed the sparse transformer.
**Aditya Ramesh** experimented with loss scaling strategies for pretraining.
**Melanie Subbiah and Arvind Neelakantan** implemented, experimented with, and tested beam search.
**Pranav Shyam** worked on SuperGLUE and assisted with connections to few-shot learning and meta-learning literature.
**Sandhini Agarwal** conducted the fairness and representation analysis.
**Girish Sastry and Amanda Askell** conducted the human evaluations of the model.
**Ariel Herbert-Voss** conducted the threat analysis of malicious use.
**Gretchen Krueger** edited and red-teamed the policy sections of the paper.
**Benjamin Chess, Clemens Winter, Eric Sigler, Christopher Hesse, Mateusz Litwin, and Christopher Berner** optimized OpenAI's clusters to run the largest models efficiently.
**Scott Gray** developed fast GPU kernels used during training.
**Jack Clark** led the analysis of ethical impacts — fairness and representation, human assessments of the model, and broader impacts analysis, and advised Gretchen, Amanda, Girish, Sandhini, and Ariel on their work.
**Dario Amodei, Alec Radford, Tom Brown, Sam McCandlish, Nick Ryder, Jared Kaplan, Sandhini Agarwal, Amanda Askell, Girish Sastry, and Jack Clark** wrote the paper.
**Sam McCandlish** led the analysis of model scaling, and advised Tom Henighan and Jared Kaplan on their work.
**Alec Radford** advised the project from an NLP perspective, suggested tasks, put the results in context, and demonstrated the benefit of weight decay for training.
**Ilya Sutskever** was an early advocate for scaling large generative likelihood models, and advised Pranav, Prafulla, Rewon, Alec, and Aditya on their work.
**Dario Amodei** designed and led the research.

**Tom Brown, Ben Mann, Prafulla Dhariwal, Dario Amodei, Nick Ryder, Daniel M Ziegler, and Jeffrey Wu**实现了大规模模型、训练基础设施和模型并行策略。
**Tom Brown、Dario Amodei、Ben Mann和Nick Ryder**进行了预训练实验。
**Ben Mann和Alec Radford**收集、过滤、去重并对训练数据进行了重叠分析。
**Melanie Subbiah、Ben Mann、Dario Amodei、Jared Kaplan、Sam McCandlish、Tom Brown、Tom Henighan和Girish Sastry**实现了下游任务和支持它们的软件框架，包括创建合成任务。
**Jared Kaplan和Sam McCandlish**最初预测巨型语言模型应该会继续获得收益，并应用缩放定律来帮助预测和指导研究中的模型和数据缩放决策。
**Ben Mann**在训练期间实现了无替换抽样。

**Alec Radford**最初证明了语言模型中存在few-shot学习。
**Jared Kaplan和Sam McCandlish**表明，在上下文中，更大的模型学习更快，并系统地研究了上下文学习曲线、任务提示和评估方法。
**Prafulla Dhariwal**实现了代码库的早期版本，并实现基于完全半精度训练的内存优化。
**Rewon Child和Mark Chen**开发了我们的模型并行策略的早期版本。
**Rewon Child和Scott Gray**贡献了稀疏transformer。
**Aditya Ramesh**尝试了预训练的损失缩放策略。
**Melanie Subbiah和Arvind Neelakantan**实现、实验和测试了波束搜索。
**Pranav Shyam**在SuperGLUE上工作，并协助与few-shot学习和元学习文献的联系。
**Sandhini Agarwal**进行了公平性和代表性分析。
**Girish Sastry和Amanda Askell**进行了模型的人类评估。
**Ariel Herbert-Voss**进行了恶意使用的威胁分析。
**Gretchen Krueger**编辑并审查了论文的政策部分。
**Benjamin Chess、Clemens Winter、Eric Sigler、Christopher Hesse、Mateusz Litwin和Christopher Berner**优化了OpenAI的集群，以高效地运行最大的模型。
**Scott Gray**开发了用于训练的快速GPU内核。
**Jack Clark**领导了对道德影响、公平性和代表性、模型的人类评估以及更广泛的影响分析的分析，并就他们的工作向Gretchen、Amanda、Girish、Sandhini和Ariel提供建议。
**Dario Amodei、Alec Radford、Tom Brown、Sam McCandlish、Nick Ryder、Jared Kaplan、Sandhini Agarwal、Amanda Askell、Girish Sastry和Jack Clark**撰写了论文。
**Sam McCandlish**领导了模型扩展的分析，并就Tom Henighan和Jared Kaplan的工作提供建议。
**Alec Radford**从NLP的角度为该项目提供了建议，提出了任务，将结果放入了上下文，并展示了权重衰减对训练的好处。
**Ilya Sutskever**是大规模生成似然模型扩展的早期倡导者，并就Pranav、Prafulla、Rewon、Alec和Aditya的工作提供建议。
**Dario Amodei**设计并领导了这项研究。

引用

**[ADG+16]** Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul,
Brendan Shillingford, and Nando De Freitas. Learning to learn by gradient descent by gradient descent.
In Advances in neural information processing systems, pages 3981–3989, 2016.
**[AI19]** WeChat AI. Tr-mt (ensemble), December 2019.
**[AJF19]** Roee Aharoni, Melvin Johnson, and Orhan Firat. Massively multilingual neural machine translation. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational
Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019.

**[BBDIW20]** Su Lin Blodgett, Solon Barocas, Hal Daume III, and Hanna Wallach. Language (technology) is power: ´ A critical survey of "bias" in nlp. arXiv preprint arXiv:2005.14050, 2020.

**[BCFL13]** Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on freebase from
question-answer pairs. In Proceedings of the 2013 conference on empirical methods in natural language
processing, pages 1533–1544, 2013.

**[BDD+09]** Luisa Bentivogli, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, and Bernardo Magnini. The fifth PASCAL recognizing textual entailment challenge. 2009.

**[BES10]** Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: an enhanced lexical
resource for sentiment analysis and opinion mining. In Lrec, volume 10, pages 2200–2204, 2010.

**[BHDD+06]** Roy Bar Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. The second PASCAL recognising textual entailment challenge. 2006.

**[BHT+20]** Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella
Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. Experience grounds language.
arXiv preprint arXiv:2004.10151, 2020.

**[BLC13]** Yoshua Bengio, Nicholas Leonard, and Aaron C. Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. Arxiv, 2013.

**[BZB+19]** Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. arXiv preprint arXiv:1911.11641, 2019.

**[Car97]** Rich Caruana. Multitask learning. Machine learning, 28(1), 1997.

**[CB78]** Susan Carey and Elsa Bartlett. Acquiring a single new word. Proceedings of the Stanford Child Language Conference, 1978.

**[CCE+18]** Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and
Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. ArXiv,
abs/1803.05457, 2018.

**[CGRS19]** Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers, 2019.

**[CHI+18]** Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke
Zettlemoyer. Quac : Question answering in context. Arxiv, 2018.
**[CLC+19]** Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina
Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. arXiv preprint
arXiv:1905.10044, 2019.
**[CLY+19]** Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and
Jingjing Liu. Uniter: Learning universal image-text representations. arXiv preprint arXiv:1909.11740,
2019.
**[Cra17]** Kate Crawford. The trouble with bias. NIPS 2017 Keynote, 2017.
**[DCLT18]** Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep
bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
**[DGM06]** Ido Dagan, Oren Glickman, and Bernardo Magnini. The PASCAL recognising textual entailment
challenge. In Machine learning challenges. evaluating predictive uncertainty, visual object classification,
and recognising textual entailment, pages 177–190. Springer, 2006.
**[DGV+18]** Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. Universal
transformers. Arxiv, 2018.
**[DHKH14]** Nadir Durrani, Barry Haddow, Philipp Koehn, and Kenneth Heafield. Edinburgh’ s phrase-based machine translation systems for wmt-14. In Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 97–104, 2014.
**[DL15]** Andrew M. Dai and Quoc V. Le. Semi-supervised sequence learning. In Advances in neural information processing systems, 2015.
**[DMST19]** Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. The CommitmentBank: Investigating projection in naturally occurring discourse. 2019. To appear in proceedings of Sinn und Bedeutung 23. Data can be found at https://github.com/mcdm/CommitmentBank/.
**[DSC+16**] Yan Duan, John Schulman, Xi Chen, Peter L. Bartlett, Ilya Sutskever, and Pieter Abbeel. Rl2
: Fast reinforcement learning via slow reinforcement learning. ArXiv, abs/1611.02779, 2016.

**[DWD+19]** Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. arXiv preprint arXiv:1903.00161, 2019.
**[DYY+19]** Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc V. Le, and Ruslan Salakhutdinov.
Transformer-xl: Attentive language models beyond a fixed-length context. Arxiv, 2019.
**[EOAG18]** Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. arXiv preprint arXiv:1808.09381, 2018.
**[FAL17]** Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of
deep networks. ArXiv, abs/1703.03400, 2017.
**[Fyo00]** Yaroslav Fyodorov. A natural logic inference system, 2000.
**[GG19]** Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. arXiv preprint arXiv:1903.03862, 2019.
**[GLT+20]** Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: Retrieval augmented language model pre-training. arXiv preprint arXiv:2002.08909, 2020.
**[GMDD07]** Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. The third PASCAL recognizing
textual entailment challenge. In Proceedings of the ACL-PASCAL workshop on textual entailment and
paraphrasing, pages 1–9. Association for Computational Linguistics, 2007.
**[Gra16]** Alex Graves. Adaptive computation time for recurrent neural networks. Arxiv, 2016.
[GSL+18] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. Annotation artifacts in natural language inference data. arXiv preprint arXiv:1803.02324, 2018.
**[GSR19]** Sebastian Gehrmann, Hendrik Strobelt, and Alexander M. Rush. Gltr: Statistical detection and visualiza tion of generated text. arXiv preprint arXiv: 1906.04043, 2019.
**[GWC+18]** Jiatao Gu, Yong Wang, Yun Chen, Kyunghyun Cho, and Victor OK Li. Meta-learning for low-resource neural machine translation. arXiv preprint arXiv:1808.08437, 2018.
**[HB20]** Daniel Hernandez and Tom Brown. Ai and efficiency, May 2020.
**[HBFC19]** Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. CoRR, abs/1904.09751, 2019.
**[HLW+20]** Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song.

Pretrained transformers improve out of distribution robustness. arXiv preprint arXiv:2004.06100, 2020.
**[HNA+17]** Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. arXiv preprint arXiv:1712.00409, 2017.
**[HR18]** Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146, 2018.
**[HVD15]** Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. arXiv
preprint arXiv:1503.02531, 2015.
**[HYC01]** Sepp Hochreiter, A Steven Younger, and Peter R Conwell. Learning to Learn Using Gradient Descent. In International Conference on Artificial Neural Networks, pages 87–94. Springer, 2001.
**[HZJ+19]** Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. Reducing sentiment bias in language models via counterfactual
evaluation. arXiv preprint arXiv:1911.03064, 2019.
**[IBGC+14]** Mohit Iyyer, Jordan Boyd-Graber, Leonardo Claudino, Richard Socher, and Hal Daume III. A neural network for factoid question answering over paragraphs. In Empirical Methods in Natural Language
Processing, 2014.
**[IDCBE19]** Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. Automatic detection of generated text is easiest when humans are fooled. arXiv preprint arXiv:1911.00650, 2019.
**[JCWZ17]** Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. arXiv preprint arXiv:1705.03551, 2017.
**[JN20]** Zheng Junyuan and Gamma Lab NYC. Numeric transformer - albert, March 2020.
**[JVS+16]** Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. Exploring the limits of language modeling. arXiv preprint arXiv:1602.02410, 2016.
**[JYS+19]** Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu.
TinyBERT: Distilling BERT for natural language understanding. arXiv preprint arXiv:1909.10351, 2019.
**[JZC+19]** Ying Ju, Fubang Zhao, Shijie Chen, Bowen Zheng, Xuefeng Yang, and Yunfeng Liu. Technical report on conversational question answering. arXiv preprint arXiv:1909.10772, 2019.

**[KCR+18]** Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL), 2018.
**[KKS+20]** Daniel Khashabi, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. Unifiedqa: Crossing format boundaries with a single qa system. arXiv preprint arXiv:2005.00700, 2020.
**[KMB20]** Sarah E. Kreps, Miles McCain, and Miles Brundage. All the news that' s fit to fabricate: Ai-generated
text as a tool of media misinformation, 2020.
**[KMH+20]** Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott
Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020.
[KPR+19] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti,
Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova,
Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural ques tions: a benchmark for question answering research. Transactions of the Association of Computational
Linguistics, 2019.
[KR16] Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. Arxiv, 2016.
[LB02] Edward Loper and Steven Bird. Nltk: The natural language toolkit, 2002.
[LC19] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. arXiv preprint
arXiv:1901.07291, 2019