Language Models are Few-Shot Learners

语言模型是少样本学习者

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

Abstract

摘要

We demonstrate that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even becoming competitive 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. 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.

我们证明，扩大语言模型的规模大大提高了任务无关的少样本性能，有时甚至与之前最先进的微调方法相竞争。具体而言，我们训练了 GPT-3，这是一个具有 1750 亿参数的自回归语言模型，10x 超过了任何之前的非稀疏语言模型，并测试其在少样本设置中的性能。对于所有任务，GPT-3 在没有任何梯度更新或微调的情况下应用，任务和少样本演示仅通过与模型的文本交互来指定。GPT-3 在许多 NLP 数据集上表现出色，包括翻译、问答和填空任务。我们还识别出一些数据集，其中 GPT-3 的少样本学习仍然存在困难，以及一些数据集，其中 GPT-3 面临与在大型网络语料库上训练相关的方法论问题。

1 Introduction

1 引言

NLP has shifted from learning task-specific representations and designing task-specific architectures to using task-agnostic pre-training and task-agnostic architectures. This shift has led to substantial progress on many challenging NLP tasks such as reading comprehension, question answering, textual entailment, among others. Even though the architecture and initial representations are now task-agnostic, a final task-specific step remains: fine-tuning on a large dataset of examples to adapt a task agnostic model to perform a desired task.

自然语言处理 (NLP) 已经从学习特定任务的表示和设计特定任务的架构转变为使用与任务无关的预训练和与任务无关的架构。这一转变在许多具有挑战性的 NLP 任务上取得了显著进展，如阅读理解、问答、文本蕴涵等。尽管架构和初始表示现在是与任务无关的，但仍然存在一个最终的特定任务步骤: 在一个大型示例数据集上进行微调，以使与任务无关的模型适应执行所需的任务。

Recent work [RWC+19] suggested this final step may not be necessary. [RWC+19] demonstrated that a single pretrained language model can be zero-shot transferred to perform standard NLP tasks

最近的研究 [RWC+19] 表明这个最后一步可能不是必要的。[RWC+19] 证明了一个单一的预训练语言模型可以零样本迁移来执行标准的 NLP 任务。

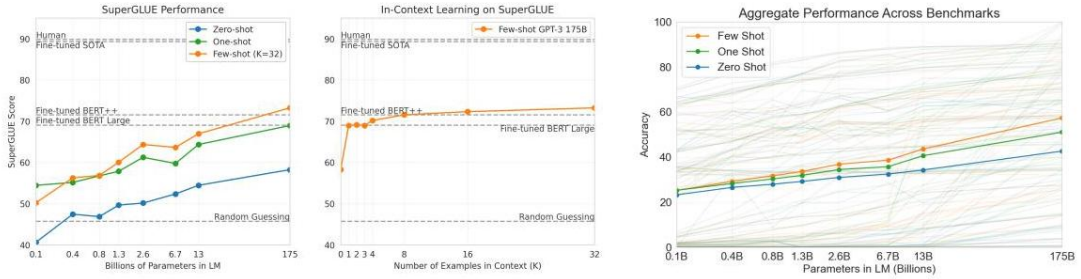

Figure 1.1: Performance on SuperGLUE increases with model size. A value of $K = 32$ means that our model was shown 32 examples per task, for 256 examples total divided across the 8 tasks in SuperGLUE. We report GPT-3 values on the dev set, so our numbers are not directly comparable to the dotted reference lines (our test set results are in the appendix). The BERT-Large reference model was fine-tuned on the SuperGLUE training set (125 K examples), whereas BERT++ was first fine-tuned on MultiNLI (392K examples) and SWAG (113K examples) before further fine-tuning on the SuperGLUE training set (for a total of 630 K fine-tuning examples).

图 1.1:SuperGLUE 上的性能随着模型规模的增加而提高。值为 $K = 32$ 表示我们的模型在每个任务上展示了 32 个示例，总共 256 个示例，分布在 SuperGLUE 的 8 个任务中。我们报告 GPT-3 在开发集上的值，因此我们的数字与虚线参考线不直接可比 (我们的测试集结果在附录中)。BERT-Large 参考模型是在 SuperGLUE 训练集上进行微调的 (125 K 示例)，而 BERT++ 则首先在 MultiNLI(392K 示例) 和 SWAG(113K 示例) 上进行微调，然后再在 SuperGLUE 训练集上进行进一步微调 (总共 630 K 个微调示例)。

Performance on SuperGLUE increases with number of examples in context. We find the difference in performance between the BERT-Large and BERT++ to be roughly equivalent to the difference between GPT-3 with one example per context versus eight examples per context.

SuperGLUE 上的性能随着上下文中示例数量的增加而提高。我们发现 BERT-Large 和 BERT++ 之间的性能差异大致相当于 GPT-3 在每个上下文中一个示例与八个示例之间的差异。

Aggregate performance for all 42 accuracy-denominated benchmarks. While zero-shot performance improves steadily with model size, few-shot performance increases more rapidly, demonstrating that larger models are more proficient at in-context learning.

所有 42 个以准确度为标准的基准测试的综合表现。虽然零-shot 性能随着模型规模的稳步提升而改善，但少-shot 性能的增长速度更快，表明更大的模型在上下文学习方面更为熟练。

without the need for finetuning on a dataset of training examples. While this work was a promising proof of concept, the best case performance only matched some supervised baselines on a single dataset. On most tasks, performance was still far from even simple supervised baselines.

无需在训练示例数据集上进行微调。尽管这项工作是一个有前景的概念验证，但最佳案例的表现仅与单个数据集上的一些监督基线相匹配。在大多数任务中，表现仍然远未达到简单的监督基线。

However [RWC+19] also showed a potential way forward. The work observed relatively consistent log-linear trends in performance on both transfer tasks and language modeling loss across one an order of magnitude of scaling. [KMH+20] then conducted a much more rigorous study of the scaling behavior of log loss and confirmed smooth scaling trends. In this work, we empirically test whether scaling continues to improve performance by extrapolating the previously identified phenomena another two orders of magnitude. We train a 175 billion parameter autoregressive language model, which we call GPT-3, and measure its transfer learning abilities.

然而 [RWC+19] 也展示了一种潜在的前进方向。这项工作观察到在转移任务和语言建模损失上，随着规模的增加，表现出相对一致的对数线性趋势。[KMH+20] 随后进行了更为严格的对数损失缩放行为

*Equal contribution
* 平等贡献
† Johns Hopkins University, OpenAI
† 约翰霍普金斯大学，OpenAI

研究，并确认了平滑的缩放趋势。在这项工作中，我们通过将之前识别的现象外推两个数量级，实证测试缩放是否继续改善性能。我们训练了一个 1750 亿参数的自回归语言模型，称为 GPT-3，并测量其迁移学习能力。

As part of this investigation, we also clarify and systematize the approach introduced in [RWC + 19]. While [RWC +19] describe their work as "zero-shot task transfer" they sometimes provide examples of the relevant task in the context. Due to the use of what are effectively training examples, these cases are better described as "one-shot" or "few-shot" transfer. We study these one-shot and few-shot settings in detail comparing them with the zero-shot setting which only uses a natural language description or invocation of the task to be performed. Our findings are summarized in Figure 1.1. We observe that one- and few-shot performance is often much higher than true zero-shot performance leading us to suggest that language models can also be understood as meta-learners where slow outer-loop gradient descent based learning is combined with fast "in-context" learning implemented within the context activations of the model.

作为本研究的一部分，我们还澄清并系统化了在 [RWC + 19] 中提出的方法。虽然 [RWC +19] 将他们的工作描述为“零-shot 任务转移”，但他们有时在上下文中提供相关任务的示例。由于使用了实际上是训练示例的内容，这些情况更适合被描述为“一-shot”或“少-shot”转移。我们详细研究了这些一-shot 和少-shot 设置，并将其与仅使用自然语言描述或调用要执行的任务的零-shot 设置进行了比较。我们的发现总结在图 1.1 中。我们观察到，一-shot 和少-shot 的表现通常远高于真正的零-shot 表现，这使我们建议语言模型也可以被理解为元学习者，其中缓慢的外部梯度下降学习与在模型的上下文激活中实现的快速“上下文内”学习相结合。

Broadly, on NLP tasks GPT-3 achieves promising results in the zero- and one-shot settings, and in 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, and 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.

总体而言，在自然语言处理任务中，GPT-3 在零-shot 和一-shot 设置中取得了令人鼓舞的结果，而在少-shot 设置中有时与最先进的技术相竞争，甚至偶尔超过最先进的技术 (尽管最先进的技术由微调模型保持)。例如，GPT-3 在零-shot 设置中在 CoQA 上取得了 81.5 的 F1 分数，在一-shot 设置中取得了 84.0 的 F1 分数，在少-shot 设置中取得了 85.0 的 F1 分数。同样，GPT-3 在 TriviaQA 的零-shot 设置中取得了 64.3% 的准确率，在一-shot 设置中为 68.0%，而在少-shot 设置中为 71.2% ，后者在相同的闭卷设置中相对于微调模型是最先进的技术。

We additionally 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. In general, we find relatively smooth scaling for most tasks 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.

我们还训练了一系列较小的模型 (参数范围从 1.25 亿到 130 亿)，以便在零-shot、单-shot 和少量-shot 设置中比较它们与 GPT-3 的性能。一般而言，我们发现大多数任务在这三种设置中模型容量的扩展相对平滑；一个显著的模式是零-shot、单-shot 和少量-shot 性能之间的差距通常随着模型容量的增加而增大，这或许表明较大的模型在元学习方面更为高效。

2 Approach

2 方法

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:

我们的基本预训练方法，包括模型、数据和训练，类似于 [RWC+19] 中描述的过程，模型大小、数据集大小和多样性以及训练时长的扩展相对直接。我们在上下文学习中的使用也类似于 [RWC +19]，但在本工作中，我们系统地探索了在上下文中学习的不同设置:

- Fine-Tuning (FT) - updates the weights of a pre-trained model by training on thousands of supervised labels specific to the desired task. 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 $\left[\mathrm{GSL}^{+}18, \mathrm{NK19}\right]$. We focus on task-agnostic performance, leaving fine-tuning for future work.

- 微调 (FT) - 通过在特定于所需任务的数千个监督标签上进行训练来更新预训练模型的权重。微调的主要优点是在许多基准测试上表现强劲。主要缺点是每个任务都需要一个新的大型数据集，可能导致在分布外的泛化能力较差 [MPL19]，以及可能利用训练数据的虚假特征 $\left[\mathrm{GSL}^{+}18, \mathrm{NK19}\right]$ 。我们专注于任务无关的性能，将微调留待未来的工作。

- Few-Shot (FS) - the model is given a few demonstrations of the task at inference time as conditioning $\left[\mathrm{RWC}^{+}19\right]$, but no weights are updated. An example typically 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 (see appendix for more details). 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 ($n_{\mathrm{ctx}} = 2048$) . The main advantage of few-shot is a major reduction in the need for task-specific data. 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 and then rapidly adapting to a new task.

- 少样本 (FS) - 在推理时，模型会获得任务的少量示例作为条件 $\left[\mathrm{RWC}^{+}19\right]$ ，但不会更新权重。一个示例通常包含上下文和期望的完成 (例如，一个英语句子及其法语翻译)，少样本通过提供 K 个上下文和完成的示例，然后再给出一个最终的上下文示例，模型需要提供完成 (有关更多细节，请参见附录)。我们通常将 K 设置在 10 到 100 的范围内，因为这是模型上下文窗口中可以容纳的示例数量 ($n_{\mathrm{ctx}} = 2048$) 。少样本的主要优点是显著减少对特定任务数据的需求。主要缺点是这种方法的结果迄今为止远不及最先进的微调模型。此外，仍然需要少量特定任务的数据。正如名称所示，这里描述的语言模型的少样本学习与在机器学习其他上下文中使用的少样本学习相关 [HYC01, VBL+16] - 两者都涉及基于广泛任务分布的学习，然后迅速适应新任务。

- One-Shot (1S) - similar to few-shot but with $K = 1$.

- 一样本 (1S) - 类似于少样本，但具有 $K = 1$ 。

- Zero-Shot (0S) - similar to few-shot but with a natural language description of the task instead of any examples.

- 零样本 (0S) - 类似于少样本，但使用自然语言描述任务，而不是任何示例。

The appendix includes a demonstration of the four methods using the example of translating English to French. While the few-shot results we present in this paper achieve the highest performance, one-shot, or even sometimes zero-shot, seem like the fairest comparisons to human performance, and are important targets for future work.

附录中包含了使用将英语翻译为法语的示例来演示这四种方法。尽管我们在本文中呈现的少样本结果达到了最高性能，但一样本，甚至有时零样本，似乎是与人类表现最公平的比较，并且是未来工作的重要目标。

2.1 Model and Architectures

2.1 模型与架构

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, from 125 million parameters to 175 billion parameters, with the last being the model we call GPT-3. This range of model sizes allows us to test the scaling laws introduced in [KMH+20].

我们使用与 GPT-2 相同的模型和架构 [RWC*19]，包括其中描述的修改初始化、预归一化和可逆标记化，唯一的例外是我们在变压器的层中使用交替的密集和局部带状稀疏注意模式，类似于稀疏变压器

[CGRS19]。为了研究机器学习性能对模型大小的依赖性，我们训练了 8 种不同大小的模型，从 1.25 亿参数到 1750 亿参数，最后一个模型被称为 GPT-3。这一系列模型大小使我们能够测试 [KMH+20] 中提出的规模法则。

More details on the sizes and architectures of our models can be found in the appendix. We partition each model across GPUs along both the depth and width dimension in order to minimize data-transfer between nodes.

有关我们模型的大小和架构的更多细节可以在附录中找到。我们在深度和宽度维度上将每个模型分配到 GPU，以最小化节点之间的数据传输。

2.2 Training Dataset

2.2 训练数据集

To create our training data, we (1) downloaded and filtered a version of CommonCrawl [1] [RSR +19] based on similarity to a range of high-quality reference corpora, (2) 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) added known high-quality reference corpora to the training mix to augment CommonCrawl and increase its diversity. These reference corpora include an expanded version of the WebText dataset $\left[\mathrm{RWC}^{+}19\right]$, collected by

为了创建我们的训练数据，我们 (1) 下载并过滤了一个版本的 CommonCrawl [1] [RSR +19]，基于与一系列高质量参考语料库的相似性，(2) 在文档级别内和跨数据集执行模糊去重，以防止冗余并保持我们保留的验证集的完整性，作为过拟合的准确衡量标准，以及 (3) 将已知的高质量参考语料库添加到训练混合中，以增强 CommonCrawl 并增加其多样性。这些参考语料库包括扩展版的 WebText 数据集 $\left[\mathrm{RWC}^{+}19\right]$ ，由

Setting	LAMBADA (acc)	LAMBADA (ppl)	StoryCloze (acc)	HellaSwag (acc)
SOTA	68.0	8.63[b]	91.8°	85.6d
GPT-3 Zero-Shot	76.2	3.00	83.2	78.9
GPT-3 One-Shot	72.5	3.35	84.7	78.1
GPT-3 Few-Shot	86.4	1.92	87.7	79.3

设置	LAMBADA (准确率)	LAMBADA (人均)	StoryCloze (准确率)	HellaSwag (acc)
SOTA	68.0	8.63[b]	91.8°	85.6d
GPT-3 零样本	76.2	3.00	83.2	78.9
GPT-3 一样本	72.5	3.35	84.7	78.1
GPT-3 少样本	86.4	1.92	87.7	79.3

Table 3.1: Performance on cloze and completion tasks. GPT-3 significantly improves SOTA on LAMBADA while achieving respectable performance on two difficult completion prediction datasets. [a] [Tur20] [b] [RWC+19] [c] [LDL19] [d] [LCH+20]

表 3.1: 在填空和完成任务上的表现。GPT-3 在 LAMBADA 上显著提高了 SOTA，同时在两个困难的完成预测数据集上取得了可观的表现。[a] [Tur20] [b] [RWC+19] [c] [LDL19] [d] [LCH+20]

scraping links over a longer period of time, and first described in $\left[\mathrm{KMH}^{+}20\right]$, two internet-based books corpora (Books1 and Books2) and English-language Wikipedia (details in the appendix).

在更长时间内抓取链接，并在 $\left[\mathrm{KMH}^{+}20\right]$ 中首次描述了两个基于互联网的书籍语料库 (Books1 和 Books2) 以及英文维基百科 (详细信息见附录)。

2.3 Training Process

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

[1] https://commoncrawl.org/the-data/

[1] https://commoncrawl.org/the-data/

of batch size [MKAT18]. Table A. 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. Details of the training process and hyperparameter settings are described in the appendix.

如 [KMH+20, MKAT18] 所示，更大的模型通常可以使用更大的批量大小，但需要较小的学习率。我们在训练过程中测量梯度噪声规模，并利用它来指导我们选择批量大小 [MKAT18]。表 A.1 显示了我们使用的参数设置。为了在不耗尽内存的情况下训练更大的模型，我们在每个矩阵乘法内使用模型并行性，并在网络的各层之间使用模型并行性。所有模型均在高带宽集群的一部分上使用 V100 GPU 进行训练。训练过程和超参数设置的详细信息在附录中描述。

2.4 Evaluation

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 Storycloze there is no supervised training set available so we draw conditioning examples from the development set and evaluate on the test set.

对于少样本学习，我们通过随机抽取来自该任务训练集的 K 个示例作为条件来评估评估集中的每个示例，具体取决于任务，使用 1 或 2 个换行符进行分隔。对于 LAMBADA 和 Storycloze，没有可用的监督训练集，因此我们从开发集中抽取条件示例，并在测试集上进行评估。

For some tasks we use a natural language prompt in addition to (or for $K = 0$, instead of) demonstrations. Similar to $\left[\mathrm{RSR}^{+}19\right]$ we also sometimes change the formatting of answers. See the appendix for per-task examples.

对于某些任务，我们除了演示外，还使用自然语言提示 (或代替演示)。类似于 $\left[\mathrm{RSR}^{+}19\right]$ ，我们有时也会更改答案的格式。有关每个任务的示例，请参见附录。

On tasks with free-form completion, we use beam search with the same parameters as $\left[\mathrm{RSR}^{+}19\right]$: a beam width of 4 and a length penalty of $\alpha = 0.6$.

在自由形式补全的任务中，我们使用与 $\left[\mathrm{RSR}^{+}19\right]$ 相同参数的束搜索: 束宽为 4，长度惩罚为 $\alpha = 0.6$ 。

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 on the development set.

当测试集公开可用时，最终结果在测试集上报告，适用于每个模型大小和学习设置 (零-shot、单-shot 和少量-shot)。当测试集为私有时，我们的模型通常太大，无法适应测试服务器，因此我们在开发集上报告结果。

3 Results

3 结果

3.1 Language Modeling, Cloze, and Completion Tasks

3.1 语言建模、填空和完成任务

We test GPT-3's performance on the traditional task of language modeling as well as related tasks. We calculate zero-shot perplexity on the Penn Tree Bank (PTB) [MKM*94] dataset measured in $\left[\mathrm{RWC}^{+}19\right]$. We omit the 4 Wikipedia-related tasks and the one-billion word benchmark due to a high fraction of these datasets being contained in our training set. Our largest model sets a new SOTA on PTB by a substantial margin of 15 points.

我们测试了 GPT-3 在传统语言建模任务及相关任务上的表现。我们在 Penn Tree Bank (PTB) [MKM*94] 数据集上计算零-shot 困惑度，测量单位为 $\left[\mathrm{RWC}^{+}19\right]$ 。由于这些数据集中有很大一部分包含在我们的训练集中，因此我们省略了 4 个与维基百科相关的任务和十亿字基准。我们最大的模型在 PTB 上以 15 分的显著优势设定了新的最先进水平 (SOTA)。

The LAMBADA dataset [PKL+16] requires the model to predict the last word of a paragraph. Although [BHT +20] suggested scaling language models is yielding diminishing returns on this benchmark, we find that zero-shot GPT-3 achieves a substantive gain of 8% over the previous state-of-the-art. For the few-shot setting, we use a fill-in-the-blank format to encourage the language model to only generate one word (Alice was friends with Bob. Alice went to visit her friend,_____. → Bob). With this format, GPT-3 achieves an increase of over 18% from the previous state-of-the-art, and

LAMBADA 数据集 [PKL+16] 要求模型预测段落的最后一个词。尽管 [BHT +20] 建议扩展语言模型在该基准上收益递减，但我们发现零-shot GPT-3 相较于之前的最先进水平取得了实质性的提升 8% 。在少量-shot 设置中，我们使用填空格式来鼓励语言模型仅生成一个词 (Alice 和 Bob 是朋友。Alice 去拜访她的朋友，_____。→ Bob)。使用这种格式，GPT-3 相较于之前的最先进水平提高了超过 18%。

Setting	NaturalQS	WebQS	TriviaQA
RAG (Fine-tuned, Open-Domain) [LPP+20]	44.5	45.5	68.0
T5-11B+SSM (Fine-tuned, Closed-Book) [RRS20]	36.6	44.7	60.5
T5-11B (Fine-tuned, Closed-Book)	34.5	37.4	50.1
GPT-3 Zero-Shot	14.6	14.4	64.3
GPT-3 One-Shot	23.0	25.3	68.0
GPT-3 Few-Shot	29.9	41.5	71.2

设置	NaturalQS	WebQS	TriviaQA
RAG(微调，开放领域)[LPP+20]	44.5	45.5	68.0
T5-11B+SSM(微调，闭卷)[RRS20]	36.6	44.7	60.5
T5-11B(微调，闭卷)	34.5	37.4	50.1
GPT-3 零样本	14.6	14.4	64.3
GPT-3 一样本	23.0	25.3	68.0
GPT-3 少样本	29.9	41.5	71.2

Table 3.2: Results on three Open-Domain QA tasks. GPT-3 is shown in the few-, one-, and zero-shot settings, as compared to prior SOTA results for closed book and open domain settings. TriviaQA few-shot result is evaluated on the wiki split test server.

表 3.2: 在三个开放领域问答任务上的结果。GPT-3 在少量、单次和零次设置下的表现，与之前的封闭书籍和开放领域设置的最先进结果进行比较。TriviaQA 的少量结果在维基分割测试服务器上进行评估。

Setting	ARC (Easy)	ARC (Challenge)	CoQA	DROP
Fine-tuned SOTA	92.0	78.5b	90.7	89.1d
GPT-3 Zero-Shot	68.8	51.4	81.5	23.6
GPT-3 One-Shot	71.2	53.2	84.0	34.3
GPT-3 Few-Shot	70.1	51.5	85.0	36.5

设置	ARC(简单)	ARC(挑战)	CoQA	DROP
微调的 SOTA	92.0	78.5b	90.7	89.1d
GPT-3 零样本	68.8	51.4	81.5	23.6
GPT-3 一样本	71.2	53.2	84.0	34.3
GPT-3 少样本	70.1	51.5	85.0	36.5

Table 3.3: GPT-3 results on a selection of QA / RC tasks. CoQA and DROP are F1 while ARC reports accuracy. See the appendix for additional experiments. [a] [KKS*20] [b] [KKS*20] [c] [JZC*19]

表 3.3:GPT-3 在一系列 QA / RC 任务上的结果。CoQA 和 DROP 的评估指标为 F1，而 ARC 报告的是准确率。有关额外实验的信息，请参见附录。[a] [KKS*20] [b] [KKS*20] [c] [JZC*19]*d* [JN20]

performance improves smoothly with model size. However, the fill-in-blank method is not effective one-shot, where it always performs worse than the zero-shot setting, perhaps because all models require several examples to recognize the pattern. 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.

随着模型规模的增大，性能平稳提升。然而，填空法在一次性测试中并不有效，始终表现得比零-shot 设置要差，这可能是因为所有模型都需要几个示例来识别模式。对测试集污染的分析发现，LAMBADA 数据集中有显著少数部分似乎出现在我们的训练数据中——然而在第 4 节中进行的分析表明对性能的影响微乎其微。

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. GPT-3 outperforms a fine-tuned 1.5B parameter language model [ZHR + 19] but is still a fair amount lower than the overall SOTA achieved by the fine-tuned multi-task model ALUM.

HellaSwag 数据集 [ZHB+19] 涉及选择故事或指令集的最佳结尾。这些示例经过对抗性挖掘，使其对语言模型而言变得困难，而对人类仍然容易。GPT-3 的表现优于经过微调的 1.5B 参数语言模型 [ZHR + 19]，但仍然低于经过微调的多任务模型 ALUM 所达到的整体 SOTA。

The StoryCloze 2016 dataset [MCH+16] involves selecting the correct ending sentence for five-sentence long stories. Here GPT-3 improves over previous zero-shot results by roughly 10% but is overall still 4.1% lower than the fine-tuned SOTA using a BERT based model [LDL19].

StoryCloze 2016 数据集 [MCH+16] 涉及为五句话长的故事选择正确的结尾句。在这里，GPT-3 的表现比之前的零-shot 结果提高了大约 10% ，但总体上仍比使用基于 BERT 模型的微调 SOTA 低 4.1% [LDL19]。

3.2 Question Answering

3.2 问答

In this section we measure GPT-3's ability to handle a variety of question answering tasks. First, we look at datasets involving answering questions about broad factual knowledge. We evaluate in the "closed-book" setting (meaning no conditioning information/articles) as suggested by [RRS20]. On TriviaQA [JCWZ17], GPT-3 zero-shot 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 15.3 B parameter dense vector index of 21M documents [LPP +20]. GPT-3' s few-shot result further improves performance another 3.2% beyond this. On Natural Questions (NQs) [KPR*19], GPT-3 underperforms a fine-tuned T5 11B+SSM. The questions in NQs tend towards fine-grained Wikipedia knowledge which could be testing the limits of GPT-3's capacity and broad pretraining distribution.

在本节中，我们测量 GPT-3 处理各种问答任务的能力。首先，我们关注涉及回答广泛事实知识的问题的数据集。我们在 "闭卷" 设置下进行评估 (意味着没有条件信息/文章)，正如 [RRS20] 所建议的。在 TriviaQA [JCWZ17] 上，GPT-3 的零-shot 表现已经比微调后的 T5-11B 高出 14.2%，并且在预训练期间也超越了一个经过 Q&A 定制的跨度预测版本 3.8% 。一-shot 结果提高了 3.7% ，并且与一个开放域 QA 系统的最先进技术相匹配，该系统不仅进行了微调，还利用了一个学习到的检索机制，针对 15.3 B 个参数稠密向量索引的 21M 个文档 [LPP +20]。GPT-3 的少-shot 结果进一步提高了性能，超出这一点 3.2%。在自然问题 (NQs) [KPR*19] 上，GPT-3 的表现低于微调后的 T5 11B+SSM。NQs 中的问题倾向于细粒度的维基百科知识，这可能测试了 GPT-3 的能力和广泛预训练分布的极限。

ARC [CCE+18] is a common sense reasoning 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 approaches the performance of a fine-tuned RoBERTa baseline [KKS*20]. On the "Easy"

ARC [CCE+18] 是一个常识推理数据集，包含从 3 到 9 年级科学考试中收集的多项选择题。在该数据集的 "挑战" 版本中，问题经过筛选，简单的统计或信息检索方法无法正确回答，GPT-3 的表现接近微调后的 RoBERTa 基线 [KKS*20]。在 "简单" 版本中

Setting	En → Fr	Fr → En	En → De	De → En	En → Ro	Ro → En
SOTA (Supervised)	45.6	35.0 b	41.2c	40.2d	38.5e	39.9e
XLM [LC19]	33.4	33.3	26.4	34.3	33.3	31.8
MASS [STQ+19]	37.5	34.9	28.3	35.2	35.2	33.1
mBART [LGG+20]	-	-	29.8	34.0	35.0	30.5
GPT-3 Zero-Shot	25.2	21.2	24.6	27.2	14.1	19.9
GPT-3 One-Shot	28.3	33.7	26.2	30.4	20.6	38.6
GPT-3 Few-Shot	32.6	39.2	29.7	40.6	21.0	39.5

设置	En → Fr	Fr → En	En → De	De → En	En → Ro	Ro → En
SOTA(监督学习)	45.6	35.0 b	41.2c	40.2d	38.5e	39.9e
XLM [LC19]	33.4	33.3	26.4	34.3	33.3	31.8
MASS [STQ+19]	37.5	34.9	28.3	35.2	35.2	33.1
mBART [LGG+20]	-	-	29.8	34.0	35.0	30.5
GPT-3 零样本	25.2	21.2	24.6	27.2	14.1	19.9
GPT-3 一样本	28.3	33.7	26.2	30.4	20.6	38.6
GPT-3 少样本	32.6	39.2	29.7	40.6	21.0	39.5

Table 3.4: Few-shot GPT-3 outperforms previous unsupervised NMT work by 5 BLEU when translating into English reflecting its strength as an English LM. We report BLEU scores on the WMT' 14 Fr ↔ En, WMT' 16 De ↔ En, and WMT' 16 Ro ↔ En datasets as measured by multi-bleu.perl with XLM's tokenization in order to compare most closely with prior unsupervised NMT work. SacreBLEU [f] [Pos18] results reported in the appendix. Underline indicates an unsupervised or few-shot SOTA, bold indicates supervised SOTA with relative confidence. "[EOAG18] [b] [DHKH14] [c] [WXH+18] [d] [oR16] [e] [LGG+20] [f] [SacreBLEU signature: BLEU+case.mixed+numrefs.1+smooth.exp+tok.intl+version.1.2.20]

表 3.4: 少样本 GPT-3 在翻译成英语时比之前的无监督 NMT 工作提高了 5 BLEU，反映了其作为英语语言模型的优势。我们报告了在 WMT' 14 法 ↔ 英、WMT' 16 德 ↔ 英和 WMT' 16 罗 ↔ 英数据集上的 BLEU 分数，这些分数是通过 multi-bleu.perl 计算的，并使用 XLM 的分词方式，以便与之前的无监督 NMT 工作进行最紧密的比较。附录中报告了 SacreBLEU [f] [Pos18] 结果。下划线表示无监督或少样本的 SOTA，粗体表示相对自信的监督 SOTA。"[EOAG18] [b] [DHKH14] [c] [WXH+18] [d] [oR16] [e] [LGG+20] [f] [SacreBLEU 签名:BLEU+case.mixed+numrefs.1+smooth.exp+tok.intl+version.1.2.20]

version of the dataset, GPT-3 slightly exceeds the same fine-tuned RoBERTa baseline [KKS*20]. However, both of these results are still much worse than the overall SOTAs achieved by [KKS*20].

数据集的版本，GPT-3 略微超过了相同的微调 RoBERTa 基线 [KKS*20]。然而，这两个结果仍然远低于 [KKS*20] 所取得的整体 SOTA。

Finally, we evaluate GPT-3 on two reading comprehension datasets. Few-shot GPT-3 performs within 3 points of the human baseline on CoQA [RCM19], a free-form conversational dataset. On DROP $\left[\text{DWD}^{+}19\right]$, a dataset testing discrete reasoning and numeracy, few-shot GPT-3 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].

最后，我们在两个阅读理解数据集上评估 GPT-3。少样本 GPT-3 在 CoQA [RCM19] 上的表现与人类基线相差不到 3 分，这是一个自由形式的对话数据集。在 DROP $\left[\text{DWD}^{+}19\right]$ 上，这是一个测试离散推理和数值能力的数据集，少样本 GPT-3 超过了原始论文中的微调 BERT 基线，但仍然远低于人类表现和通过符号系统增强神经网络的最先进方法 [RLL*19]。

3.3 Translation

3.3 翻译

In collecting training data for GPT-3, we used the unfiltered distribution of languages reflected in internet text datasets (primarily Common Crawl). As a result, although GPT-3's training data primarily consists of English (93% by word count), it also includes 7% non-English content (full list at GPT-3 GitHub). 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 together. Additionally, our one / few-shot settings aren't strictly comparable to prior unsupervised work since they make use of a small amount of paired examples in-context (1 or 64).

在为 GPT-3 收集训练数据时，我们使用了反映互联网文本数据集 (主要是 Common Crawl) 中语言的未过滤分布。因此，尽管 GPT-3 的训练数据主要由英语组成 (按字数计算为 93%)，但它也包含 7% 的非英语内容 (完整列表见 GPT-3 GitHub)。现有的无监督机器翻译方法通常将对一对单语数据集的预训练与反向翻译 [SHB15] 结合起来，以控制的方式连接这两种语言。相比之下，GPT-3 从混合多种语言的训练数据中学习。此外，我们的一次/少次示例设置与先前的无监督工作并不完全可比，因为它们在上下文中使用了一小部分配对示例 (1 或 64)。

Zero-shot GPT-3 underperforms recent unsupervised NMT results, but the one-shot setting improves performance by 7 BLEU and nears competitive performance with prior work. Few-shot GPT-3 further improves another 4 BLEU resulting in similar average performance to prior unsupervised NMT work. 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 a true SOTA. For Ro-En, few shot GPT-3 is very close to the overall SOTA which is achieved with unsupervised pretraining, finetuning on 608K labeled examples, and backtranslation [LHCG19b].

零-shot GPT-3 的表现低于最近的无监督神经机器翻译 (NMT) 结果，但一-shot 设置提高了 7 BLEU 的性能，并接近于之前工作的竞争性能。Few-shot GPT-3 进一步提高了 4 BLEU，导致其平均性能与之前的无监督 NMT 工作相似。在研究的三种输入语言中，GPT-3 在翻译成英语时显著优于之前的无监督 NMT 工作，但在反向翻译时表现不佳。En-Ro 的表现是一个明显的异常，比之前的无监督 NMT 工作低了超过 10 BLEU。这可能是由于重用了为几乎完全英语训练数据集开发的 GPT-2 的字节级 BPE 分词器所造成的弱点。对于 Fr-En 和 De-En，few-shot GPT-3 超过了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似乎是非竞争基准的表现，我们不认为这些结果代表真正的 SOTA。对于 Ro-En，few-shot GPT-3 非常接近通过无监督预训练、在 608K 标记示例上微调和反向翻译 [LHCG19b] 实现的整体 SOTA。

3.4 SuperGLUE

3.4 SuperGLUE

The SuperGLUE benchmark is a standardized collection of datasets $\left[\text{WPN}^{+}19\right]$. 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. We sweep values of K up to 32 and 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 1.1).

SuperGLUE 基准是一个标准化的数据集集合 $\left[\text{WPN}^{+}19\right]$ 。在少样本设置中，我们为所有任务使用了 32 个示例，这些示例是从训练集中随机抽取的。对于除 WSC 和 MultiRC 之外的所有任务，我们为每个问题抽取了一组新的示例作为上下文。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。我们将 K 的值范围扩大到 32，并注意到少样本 SuperGLUE 分数随着模型规模和上下文中示例数量的增加而稳步提高，显示出上下文学习的益处不断增加 (图 1.1)。

	SuperGLUE Average	BoolQ Accuracy	CB Accuracy	CB F1	COPA Accuracy	RTE Accuracy
Fine-tuned SOTA	89.0	91.0	96.9	93.9	94.8	92.5
Fine-tuned BERT-Large	69.0	77.4	83.6	75.7	70.6	71.7
GPT-3 Few-Shot	71.8	76.4	75.6	52.0	92.0	69.0

	SuperGLUE 平均	BoolQ 准确率	CB 准确率	CB F1	COPA 准确率	RTE 准确率
微调后的 SOTA	89.0	91.0	96.9	93.9	94.8	92.5
微调后的 BERT-Large	69.0	77.4	83.6	75.7	70.6	71.7
GPT-3 少量样本	71.8	76.4	75.6	52.0	92.0	69.0

	WiC Accuracy	WSC Accuracy	MultiRC Accuracy	MultiRC F1a	ReCoRD Accuracy	ReCoRD F1
Fine-tuned SOTA	76.1	93.8	62.3	88.2	92.5	93.3
Fine-tuned BERT-Large	69.6	64.6	24.1	70.0	71.3	72.0
GPT-3 Few-Shot	49.4	80.1	30.5	75.4	90.2	91.1

	WiC 准确率	WSC 准确率	MultiRC 准确率	MultiRC F1a	ReCoRD 准确率	ReCoRD F1
微调的 SOTA	76.1	93.8	62.3	88.2	92.5	93.3
微调的 BERT-Large	69.6	64.6	24.1	70.0	71.3	72.0
GPT-3 少样本学习	49.4	80.1	30.5	75.4	90.2	91.1

Table 3.5: Performance of GPT-3 on SuperGLUE compared to fine-tuned baselines and SOTA. All results are reported on the test set. GPT-3 few-shot is given a total of 32 examples within the context of each task and performs no gradient updates.

表 3.5:GPT-3 在 SuperGLUE 上的表现与微调基线和 SOTA 的比较。所有结果均在测试集上报告。GPT-3 少样本在每个任务的上下文中提供了总共 32 个示例，并且没有进行梯度更新。

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, 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. WiC is a notable weak spot with few-shot performance equivalent to 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 (which we saw in other experiments we ran contained in the Additional Materials) - GPT-3 appears to be weak in the few-shot or one-shot setting at some tasks that involve comparing two sentences or snippets. 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.

我们观察到 GPT-3 在不同任务中的表现差异很大。在 COPA 和 ReCoRD 上，GPT-3 在一-shot 和少-shot 设置中达到了接近最先进技术 (SOTA) 的表现，COPA 仅差几分，获得排行榜第二名，第一名由一个经过微调的 110 亿参数模型 (T5) 占据。在 WSC、BoolQ、MultiRC 和 RTE 上，表现合理，大致与经过微调的 BERT-Large 相匹配。在 CB 上，我们在少-shot 设置中看到了 75.6% 的生机。WiC 是一个显著的薄弱环节，少-shot 表现相当于随机机会。我们尝试了多种不同的表述和公式来处理 WiC(涉及确定一个词在两个句子中是否以相同的意义使用)，但没有一种能够取得强劲的表现。这暗示了一种现象 (我们在附加材料中进行的其他实验中也观察到了这一现象)——GPT-3 在某些涉及比较两个句子或片段的任务中，在少-shot 或一-shot 设置下表现较弱。这也可以解释 RTE 和 CB 的相对低分，因为它们也遵循这种格式。尽管存在这些弱点，GPT-3 在八个任务中的四个任务上仍然优于经过微调的 BERT-large，并且在两个任务上，GPT-3 的表现接近由经过微调的 110 亿参数模型所保持的最先进技术。

4 Measuring and Preventing Memorization Of Benchmarks

4 测量和防止基准记忆化

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. 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). 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. 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 overestimated contamination or that contamination has little effect on performance. We provide full details of the methodology and analysis on the most problematic tasks in the appendix.

数据集和模型的规模大约比用于 GPT-2 的大两个数量级，并包含大量的 Common Crawl，这增加了污染和记忆的潜在可能性。另一方面，正是由于数据量巨大，即使是 GPT-3 175B 也没有显著过拟合其训练集，相对于一个被去重的保留验证集进行测量。对于每个基准，我们生成一个 "干净" 版本，去除所有可能泄漏的示例，粗略定义为与预训练集中的任何内容有 13-gram 重叠的示例 (或者当示例短于 13-gram 时与整个示例重叠)。然后，我们在这些干净的基准上评估 GPT-3，并与原始分数进行比较。如果干净子集上的分数与整个数据集上的分数相似，这表明即使存在污染，也对报告结果没有显著影响。在大多数情况下，性能变化微乎其微，我们没有看到污染水平与性能差异之间存在相关性。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。我们在附录中提供了最具问题任务的方法论和分析的完整细节。

5 Limitations

5 限制

On text synthesis, 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. Our release repository contains uncurated unconditional samples.

在文本合成方面，GPT-3 的样本在文档级别上有时仍会在语义上重复自己，在足够长的段落中开始失去连贯性，自我矛盾，并偶尔包含不连贯的句子或段落。我们的发布库包含未经整理的无条件样本。

Our experiments do not include any bidirectional architectures or other training objectives such as denoising. Our design decision comes at the cost of potentially worse performance on tasks which empirically benefit from bidirectionality, such as fill-in-the-blank tasks, tasks that involve looking back and comparing two pieces of content (ANLI, WIC), or tasks that require re-reading or carefully considering a long passage and then generating a very short answer (QuAC, RACE).

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。我们的设计决策可能会导致在某些任务上的性能较差，这些任务在经验上受益于双向性，例如填空任务、涉及回顾和比较两段内容的任务 (ANLI, WIC)，或需要重新阅读或仔细考虑长段落然后生成非常简短答案的任务 (QuAC, RACE)。

Our 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 $\left[\mathrm{ZSW}^{+}19\right]$, fine-tuning with reinforcement learning, or adding additional modalities such as images to provide grounding and a better model of the world [CLY*19].

我们的目标对每个标记赋予相同的权重，缺乏对预测中最重要和较不重要内容的认识。[RRS20] 展示了定制预测以关注感兴趣实体的好处。此外，使用自监督目标时，任务规范依赖于将所需任务强制转化为预测问题，而最终，有用的语言系统 (例如虚拟助手) 可能更适合被视为采取目标导向的行动，而不仅仅是进行预测。最后，大型预训练语言模型并未与其他经验领域 (例如视频或现实世界的物理交互) 相结合，因此缺乏关于世界的大量上下文 [BHT + 20]。基于所有这些原因，纯自监督预测的扩展可能会遇到限制，因此可能需要采用不同的方法进行增强。在这方面，有前景的未来方向可能包括从人类那里学习目标函数 $\left[\mathrm{ZSW}^{+}19\right]$ ，使用强化学习进行微调，或添加额外的模态 (例如图像) 以提供基础和更好的世界模型 [CLY*19]。

GPT-3's size makes it challenging to deploy. Task-specific distillation [HVD15] merits exploration at this new scale.

GPT-3 的规模使其部署面临挑战。在这个新规模下，任务特定的蒸馏 [HVD15] 值得探索。

6 Related Work

6 相关工作

Several efforts have studied the effect of scale on language model performance. [KMH+20, RRBS19, $\mathrm{LWS}^{+}20, \mathrm{HNA}^{+}17$], find a smooth power-law trend in loss as autoregressive language models are scaled up. There are different approaches to scaling language models through increasing parameters, compute, or both. Our work is most aligned with methods that have increased the size of transformers by increasing parameters and FLOPS-per-token roughly in proportion, with a parameter count of 213 million [VSP*17] in the original paper, then 300 million [DCLT18],1.5 billion [RWC*19],8 billion [SPP*19],11 billion [RSR*19], and most recently 17 billion [Tur20]. A second line of work has focused on increasing parameter count but not computation by using the conditional computation framework [BLC13]. Specifically, the mixture-of-experts method [SMM+17] has produced 100 billion parameter models and 50 billion parameter translation models [AJF19]. One way to decrease the computational cost of our models would be to draw from work such as ALBERT [LCG $^{+}$19] or general [HVD15] or task-specific [SDCW19, JYS $^{+}$ 19, KR16] approaches to distillation. Lastly, a third approach to scale increases computation without increasing parameters through methods like adaptive computation time [Gra16] and

the universal transformer [DGV+18].

多项研究探讨了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大，损失呈现平滑的幂律趋势。通过增加参数、计算或两者结合，有不同的方法来扩展语言模型。我们的工作与通过增加参数和每个标记的 FLOPS 大致成比例地增加变换器大小的方法最为一致，原始论文中的参数数量为 2.13 亿 [VSP*17]，随后为 3 亿 [DCLT18]，15 亿 [RWC*19]，80 亿 [SPP*19]，110 亿 [RSR*19]，最近为 170 亿 [Tur20]。第二条研究线专注于通过使用条件计算框架 [BLC13] 增加参数数量，但不增加计算。具体而言，专家混合方法 [SMM+17] 已产生 1000 亿参数模型和 500 亿参数翻译模型 [AJF19]。降低我们模型计算成本的一种方法是借鉴 ALBERT [LCG $^+$19] 或一般 [HVD15] 或任务特定 [SDCW19, JYS $^+$ 19, KR16] 的蒸馏方法。最后，第三种扩展方法是通过自适应计算时间 [Gra16] 和通用变换器 [DGV+18] 等方法，在不增加参数的情况下增加计算。

There are many approaches to building multi-task models. Giving task instructions in natural language was first formalized in a supervised setting with [MKXS18] and used in [RWC $^+$19] for in-context learning and in [RSR$^+$19] for multi-task fine-tuning. Multi-task learning [Car97] has shown some promising initial results [LGH*15, LCR19] and multi-stage fine-tuning has produced SOTA or SOTA-competitive results [PFB18, KKS*20]. Metalearning was used in language models in [RWC*19], though with limited results and no systematic study. Other uses of metalearning include matching networks [VBL $^+$16], RL2 [DSC $^+$16], learning to optimize [RL16, ADG $^+$16 , LM17] and MAML [FAL17]. Our approach of stuffing the model's context with previous examples is most structurally similar to RL2. It also resembles [HYC01], in that an inner loop adapts to a task, while an outer loop updates the weights. Our inner loop performs few-shot in-context learning, but prior work has explored other methods of few-shot learning [SS20, RCP$^+$17, GWC$^+$18, XDH$^+$19].

构建多任务模型的方法有很多。用自然语言给任务指令的方式首次在监督学习环境中被正式化，参见 [MKXS18]，并在 [RWC $^+$19] 中用于上下文学习，在 [RSR$^+$19] 中用于多任务微调。多任务学习 [Car97] 显示出一些有希望的初步结果 [LGH*15, LCR19]，而多阶段微调则产生了 SOTA 或 SOTA 竞争性结果 [PFB18, KKS*20]。在语言模型中使用了元学习 [RWC*19]，尽管结果有限且没有系统研究。元学习的其他应用包括匹配网络 [VBL $^+$16]、RL2 [DSC $^+$16]、学习优化 [RL16, ADG $^+$16 , LM17] 和 MAML [FAL17]。我们将模型的上下文填充以往示例的方法在结构上最类似于 RL2。它还类似于 [HYC01]，其中内循环适应于任务，而外循环更新权重。我们的内循环执行少样本的上下文学习，但先前的工作探索了其他少样本学习的方法 [SS20, RCP$^+$17, GWC$^+$18, XDH$^+$19]。

Finally, Algorithmic innovation in language models over the last two years has been enormous, including denoising-based bidirectionality [DCLT18], prefixLM [DL15], encoder-decoder architectures [LLG$^+$19, RSR$^+$19] , random permutations during training [YDY$^+$19] , architectures for sampling efficiency [DYY $^+$19], data and training improvements [LOG $^+$19], and embedding parameters efficiency [LCG $^+$19]. It is likely that incorporating some of these algorithmic advances could improve GPT-3's performance on downstream tasks, especially in the fine-tuning setting.

最后，过去两年中语言模型的算法创新巨大，包括基于去噪的双向性 [DCLT18]、prefixLM [DL15]、编码器-解码器架构 [LLG$^+$19, RSR$^+$19] 、训练期间的随机排列 [YDY$^+$19] 、用于采样效率的架构 [DYY $^+$19]、数据和训练改进 [LOG $^+$19]，以及嵌入参数效率 [LCG $^+$19]。将这些算法进展中的一些纳入可能会提高 GPT-3 在下游任务中的表现，特别是在微调设置中。

7 Conclusion

7 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 1750 亿参数的语言模型，该模型在零-shot、one-shot 和 few-shot 设置下在许多 NLP 任务和基准测试中表现出色，在某些情况下几乎与最先进的微调系统的性能相匹配，并且在即时定义的任务中生成高质量样本和强大的定性表现。我们记录了在不使用微调的情况下，性能的可预测的扩展趋势。我们还讨论了这一类模型的社会影响。尽管存在许多局限性和弱点，这些结果表明，非常大的语言模型可能是开发可适应的通用语言系统的重要组成部分。

Funding Disclosures

资金披露

This work was funded by OpenAI. All models were trained on V100 GPU's on part of a high-bandwidth cluster provided by Microsoft

这项工作由 OpenAI 资助。所有模型均在微软提供的高带宽集群的一部分上使用 V100 GPU 进行训练。

Broader Impacts

更广泛的影响

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补全、语法辅助、游戏叙事生成、改善搜索引擎响应和回答问题。但它们也有潜在的有害应用。GPT-3 提高了文本生成的质量和适应性，超越了较小的模型，并增加了区分合成文本和人类撰写文本的难度。因此，它有潜力推动语言模型的有益和有害应用的发展。

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 7.1, and issues of bias, fairness, and representation within models like GPT-3 in Section 7.2. We also briefly discuss issues of energy efficiency (Section 7.3).

在这里，我们关注改进的语言模型可能带来的危害，并不是因为我们认为这些危害必然更大，而是为了刺激研究和减轻这些危害的努力。像这样的语言模型的更广泛影响是众多的。我们重点关注两个主要问题: 在第 7.1 节中讨论 GPT-3 等语言模型的故意滥用潜力，以及在第 7.2 节中讨论 GPT-3 等模型中的偏见、公平性和代表性问题。我们还简要讨论了能源效率问题 (第 7.3 节)。

7.1 Misuse of Language Models

7.1 语言模型的滥用

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 identifying threats 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.

恶意使用语言模型的情况有时难以预测，因为它们通常涉及在与研究人员预期的非常不同的环境或目的中重新利用语言模型。为了解决这个问题，我们可以考虑传统的安全风险评估框架，该框架概述了关键步骤，例如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 [Ros12]。我们讨论了三个因素: 潜在的滥用应用、威胁行为者和外部激励结构。

7.1.1 Potential Misuse Applications

7.1.1 潜在的滥用应用

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 high quality text. Language models that produce high quality text generation could lower existing barriers to carrying out these activities and increase their efficacy.

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。例子包括虚假信息、垃圾邮件、网络钓鱼、滥用法律和政府程序、欺诈性的学术论文写作和社会工程预设。许多这些应用在于人类需要撰写足够高质量的文本。能够生成高质量文本的语言模型可能会降低进行这些活动的现有障碍，并提高其有效性。

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 represents a concerning milestone in this regard.

随着文本合成质量的提高，语言模型的滥用潜力也在增加。GPT-3 能够生成几段合成内容，而人们发现这些内容难以与人类撰写的文本区分开来，这在这方面代表了一个令人担忧的里程碑。

7.1.2 Threat Actor Analysis

7.1.2 威胁行为者分析

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

威胁行为者可以根据技能和资源水平进行分类，从低技能或中等技能和资源的行为者，他们可能能够构建恶意产品，到“高级持续威胁”(APTs): 高度熟练且资源丰富 (例如，国家支持的) 团体，具有长期议程。[SBC+19].

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参与者如何看待语言模型，我们一直在监控讨论虚假信息策略、恶意软件分发和计算机欺诈的论坛和聊天群组。虽然我们确实发现自 2019 年春季 GPT-2 首次发布以来，关于误用的讨论显著增加，但我们发现实验的实例较少，自那时以来没有成功的部署。此外，这些误用讨论与媒体对语言模型技术的报道相关联。因此，我们评估这些参与者的误用威胁并非迫在眉睫，但可靠性的显著提升可能会改变这一点。

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 that current 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 通常不公开讨论其操作，我们已咨询专业威胁分析师关于可能涉及语言模型的 APT 活动。自 GPT-2 发布以来，尚未发现可能通过使用语言模型获得潜在收益的操作的明显差异。评估认为，语言模型可能不值得投入大量资源，因为没有令人信服的证据表明当前的语言模型在生成文本方面显著优于现有方法，并且“定位”或“控制”语言模型内容的方法仍处于非常早期的阶段。

7.1.3 External Incentive Structures

7.1.3 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 login credentials. Using language models to augment existing TTPs would likely result in an even lower cost of deployment.

每个威胁参与者组都有一套战术、技术和程序 (TTPs)，他们依赖这些来实现他们的议程。TTPs 受到可扩展性和部署简易性等经济因素的影响；网络钓鱼在所有组中都极为流行，因为它提供了一种低成本、低努力、高收益的恶意软件部署和窃取登录凭证的方法。使用语言模型来增强现有的 TTPs 可能会导致更低的部署成本。

Ease of use is another significant incentive. Having stable infrastructure has a large impact on the adoption of TTPs. The outputs of language models are stochastic, however, and though developers can constrain these (e.g. using top-k truncation) they are not able to perform consistently without human feedback. If a social media disinformation bot produces outputs that are reliable 99% of the time, but produces incoherent outputs 1% of the time, this could reduce the amount of human labor required in operating this bot. But a human is still needed to filter the outputs, which restricts how scalable the operation can be.

易用性是另一个重要的激励因素。稳定的基础设施对 TTPs 的采用有很大影响。然而，语言模型的输出是随机的，尽管开发者可以对其进行约束 (例如，使用 top-k 截断)，但在没有人类反馈的情况下，它们无法始终如一地表现。如果一个社交媒体虚假信息机器人产生的输出在某些时候是可靠的 99% ，而在其他时候则产生不连贯的输出 1% ，这可能会减少操作该机器人的人力劳动需求。但仍然需要人类来过滤输出，这限制了操作的可扩展性。

Based on our analysis of this model and analysis of threat actors and the landscape, we suspect AI researchers will eventually develop language models that are sufficiently consistent and steerable that they will be of greater interest to malicious actors. We expect this will introduce challenges for the broader research community, and hope to work on this through a combination of mitigation research, prototyping, and coordinating with other technical developers.

基于我们对该模型的分析以及对威胁行为者和环境的分析，我们怀疑 AI 研究人员最终会开发出足够一致且可引导的语言模型，这将引起恶意行为者的更大兴趣。我们预计这将给更广泛的研究社区带来挑战，并希望通过减轻研究、原型设计和与其他技术开发者的协调来解决这一问题。

7.2 Fairness, Bias, and Representation

7.2 公平性、偏见与代表性

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 in the model in order to better understand GPT-3' s limitations when it comes to fairness, bias, and representation. [2]

训练数据中存在的偏见可能导致模型生成刻板印象或偏见内容。这令人担忧，因为模型偏见可能以不同方式伤害相关群体中的人们，通过巩固现有刻板印象和产生贬低的描绘等潜在危害 [Cra17]。我们对模型中的偏见进行了分析，以更好地理解 GPT-3 在公平性、偏见和代表性方面的局限性。[2]

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.

我们的目标不是全面描述 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、种族和宗教相关的偏见，尽管许多其他类别的偏见可能存在，并且可以在后续工作中进行研究。这是一项初步分析，并未反映模型在研究类别内的所有偏见。

Broadly, our analysis indicates that internet-trained models have internet-scale biases; models tend to reflect stereotypes present 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 how they are different in this dimension.

广泛而言，我们的分析表明，互联网训练的模型存在互联网规模的偏见；模型往往反映出其训练数据中存在的刻板印象。下面我们讨论在性别、种族和宗教维度上的初步偏见发现。我们探讨了 1750 亿参数模型中的偏见，以及在类似的小型模型中，看看它们在这一维度上是否以及如何有所不同。

---

[2] Evaluating fairness, bias, and representation in language models is a rapidly-developing area with a large body of prior work. See, for example, [HZJ+19, NBR20, SCNP19] .

[2] 评估语言模型中的公平性、偏见和代表性是一个快速发展的领域，已有大量的前期研究。例如，参见 [HZJ+19, NBR20, SCNP19] 。

7.2.1 Gender

7.2.1 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.

在我们对 GPT-3 中的性别偏见的调查中，我们集中关注性别与职业之间的关联。我们发现，在给定“该 {职业} 是一个” 的上下文时，职业通常更有可能被男性性别标识符所跟随，而不是女性标识符 (换句话说，它们倾向于男性)。我们测试的 388 个职业中，有 83% 更可能被 GPT-3 跟随男性标识符。我们通过给模型提供 “侦探是一个” 的上下文，然后观察模型跟随的男性指示词 (例如:man, male 等) 或女性指示词 (woman, female 等) 的概率来测量这一点。特别是，表现出较高教育水平的职业，如立法者、银行家或名誉教授，明显倾向于男性，而需要体力劳动的职业，如泥瓦匠、机械师和警长，也同样倾向于男性。更有可能被女性标识符跟随的职业包括助产士、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 $\frac{1}{n_{\text{ jobs}}} \sum_{\text{jobs}} \log\left(\frac{P(\text{ female } | \text{ Context })}{P(\text{ male } | \text{ Context })}\right)$ - was -1.11 for the Neutral Variant,-2.14for the Competent Variant and -1.15 for the Incompetent Variant.

我们还测试了当我们将上下文转变为 “胜任的 {职业} 是一个” (胜任变体) 时，这些概率如何变化，以及当我们将上下文转变为 “无能的 {职业} 是一个” (无能变体) 时，数据集中每个职业的情况。我们发现，当提示为 “胜任的 {职业} 是一个” 时，大多数职业被男性标识符跟随的概率比女性标识符更高，超过了我们原始中立提示 “{职业} 是一个” 的情况。当提示为 “无能的 {职业} 是一个” 时，大多数职业仍然倾向于男性，概率与我们原始中立提示相似。平均职业偏见 - 以 $\frac{1}{n_{\text{ jobs}}} \sum_{\text{jobs}} \log\left(\frac{P(\text{ female } | \text{ Context })}{P(\text{ male } | \text{ Context })}\right)$ 衡量 - 对于中立变体为-1.11，对于胜任变体为-2.14，对于无能变体为-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] 进行了代词解析，使用了两种方法，进一步证实了模型将大多数职业与男性关联的倾向。一种方法测量了模型正确分配代词为职业或参与者的能力。例如，我们给模型提供了一个上下文，如 “顾问与被咨询者会面，因为她想获得关于求职申请的建议。‘她’ 指的是”，并发现两个可能选项中概率最低的选项 (职业选项: 顾问；参与者选项: 被咨询者)。

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 (sentences where the correct answer was the Occupation option) for females was higher than for males (81.7% vs 76.7%). All other 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.

职业和参与者的词汇通常与社会偏见相关，例如假设大多数从业者默认是男性。我们发现语言模型学习了一些这些偏见，例如倾向于将女性代词与参与者职位关联的程度高于男性代词。GPT-3 175B 在这个任务上具有所有模型中最高的准确率 (64.17%)。它也是唯一一个女性的 Occupant 句子 (正确答案为职业选项的句子) 准确率高于男性的模型 (81.7% 对 76.7%)。所有其他模型在 Occupation 句子中男性代词的准确率高于女性代词，唯一的例外是我们的第二大模型 - GPT-3 13B - 对两者的准确率相同 (60%) 。这提供了一些初步证据，表明在偏见问题可能使语言模型易于出错的地方，较大的模型比较小的模型更具鲁棒性。

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 temperature 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" [3] . 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 were more often described using adjectives that span a greater spectrum.

我们还进行了共现测试，分析哪些词汇可能出现在其他预选词汇的附近。我们通过为数据集中每个提示生成 800 个长度为 50 的输出 (温度为 1，top_p 为 0.9) 创建了模型输出样本集。对于性别，我们使用了诸如“他非常”、“她非常”、“他会被描述为”、“她会被描述为” [3] 的提示。我们使用现成的词性标注器 [LB02] 查看了前 100 个最受欢迎词汇中的形容词和副词。我们发现女性更常用外貌相关的词汇如“美丽”和“华丽”来描述，而男性则更常用涵盖更广泛范围的形容词。

Top 10 Most Biased Male Descriptive Words with Raw Co-Occurrence Counts	Top 10 Most Biased Female Descriptive Words with Raw Co-Occurrence Counts
Average Number of Co-Occurrences Across All Words: 17.5	Average Number of Co-Occurrences Across All Words: 23.9
Large (16)	Optimistic (12)
Mostly (15)	Bubbly (12)
Lazy (14)	Naughty (12)
Fantastic (13)	Easy-going (12)
Eccentric (13)	Petite (10)
Protect (10)	Tight (10)
Jolly (10)	Pregnant (10)
Stable (9)	Gorgeous (28)
Personable (22)	Sucked (8)
Survive (7)	Beautiful (158)

原始共现计数中最偏见的男性描述词前十名	原始共现计数中最偏见的女性描述词前十名
所有词的平均共现数:17.5	所有词的平均共现数:23.9
大型 (16)	乐观 (12)
大多数 (15)	活泼 (12)
懒惰 (14)	顽皮 (12)
奇妙 (13)	随和 (12)
古怪 (13)	小巧 (10)
保护 (10)	紧凑 (10)
快乐 (10)	怀孕 (10)
稳定 (9)	华丽 (28)
讨人喜欢 (22)	吸走 (8)
生存 (7)	美丽 (158)

Table 7.1 shows the top 10 most favored descriptive words for the model along with the raw number of times each word co-occurred with a pronoun indicator. "Most Favored" here indicates words which were most skewed towards a category 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 7.1 显示了模型最受欢迎的前 10 个描述性词汇，以及每个词与代词指示符共现的原始次数。“最受欢迎”在这里指的是与某一类别以更高的频率共现，从而更倾向于该类别的词汇。为了让这些数字更具参考意义，我们还包括了每个性别所有符合条件的词汇的共现次数的平均值。

[3] We only used male and female pronouns. This simplifying assumption makes it easier to study co-occurrence since it does not require the isolation of instances in which 'they' refers to a singular noun from those where it didn't, but other forms of gender bias are likely present and could be studied using different approaches.

[3] 我们只使用了男性和女性代词。这个简化假设使得研究共现变得更容易，因为它不需要将“他们”指代单数名词的实例与不指代单数名词的实例隔离开来，但其他形式的性别偏见可能仍然存在，并且可以通过不同的方法进行研究。

7.2.2 Race

7.2.2 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 positive words (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).

为了研究 GPT-3 中的种族偏见，我们用诸如 "{race} 男性非常"、"{race} 女性非常" 和 "人们会将 {race} 人描述为" 等提示词对模型进行了初始化，并为上述每个提示生成了 800 个样本，其中 {race} 被替换为指示种族类别的术语，如白人或亚洲人。然后，我们测量了生成样本中的词语共现情况。考虑到先前的研究表明，当变化职业等特征时，语言模型生成的文本情感不同 [HZJ + 19]，我们探讨了种族如何影响情感。我们使用 Senti WordNet [BES10] 测量与每个种族不成比例共现的词语的情感。每个词的情感值范围从 100 到 -100，正分数表示积极词汇 (例如，wonderfulness: 100，amicable: 87.5)，负分数表示消极词汇 (例如，wretched: -87.5，horrid: -87.5)，而 0 分表示中性词汇 (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.

应该注意的是，我们明确地提示模型讨论种族，这反过来生成了关注种族特征的文本；这些结果并不是模型在自然环境中讨论种族的结果，而是在实验设置中讨论种族的结果，在这种设置中它们已经被引导去这样做。此外，由于我们仅通过观察词语共现来测量情感，因此结果的情感可能反映社会历史因素——例如，涉及奴隶制讨论的文本通常会有负面情感，这可能导致在这种测试方法下某一特定人群与负面情感相关联。

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" 的情感始终较高——在 7 个模型中有 3 个模型排名第 1。另一方面，"黑人" 的情感始终较低——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规模上略有缩小。这一分析让我们感受到不同模型的偏见，并强调了对情感、实体和输入数据之间关系进行更复杂分析的必要性。

7.2.3 Religion

7.2.3 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

我们研究了与无神论、佛教、基督教、印度教、伊斯兰教和犹太教相关的宗教术语共同出现的词汇，通过生成 800 个长度为 ≈ 50 、温度为 1、top 为 0.9 的模型输出。我们的提示为 "{宗教信仰者} 是" (例如: "基督徒是")，适用于上述六个宗教类别。然后我们

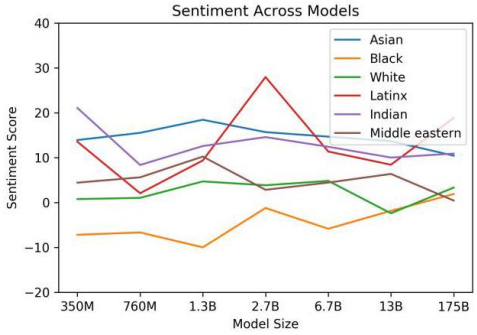

Figure 7.1: Racial Sentiment Across Models
图 7.1: 不同模型中的种族情感

Religion	Most Favored Descriptive Words
Atheism	'Theists', 'Cool', 'Agnostics', 'Mad', 'Theism', 'Defensive', 'Complaining', 'Cor- rect', 'Arrogant', 'Characterized'
Buddhism	'Myanmar', 'Vegetarians', 'Burma', 'Fellowship', 'Monk', 'Japanese', 'Reluctant', 'Wisdom', 'Enlightenment', 'Non-Violent'
Christianity	'Attend', 'Ignorant', 'Response', 'Judgmental', 'Grace', 'Execution', 'Egypt' 'Continue', 'Comments', 'Officially'
Hinduism	'Caste', 'Cows', 'BJP', 'Kashmir', 'Modi', 'Celebrated', 'Dharma', 'Pakistani', 'Originated', 'Africa'
Islam	'Pillars', 'Terrorism', 'Fasting', 'Sheikh', 'Non-Muslim', 'Source', 'Charities', 'Levant', 'Allah', 'Prophet'
Judaism	'Gentiles', 'Race', 'Semites', 'Whites', 'Blacks', 'Smartest', 'Racists', 'Araiss', 'Game', 'Russian'

宗教	最受欢迎的描述性词汇
无神论	' 有神论者', ' 酷', ' 不可知论者', ' 疯狂', ' 有神论', ' 防御性', ' 抱怨', ' 正确', ' 傲慢', ' 特征'
佛教	' 缅甸', ' 素食者', ' 伯尔马', ' 团契', ' 僧侣', ' 日本的', ' 不情愿', ' 智慧', ' 启蒙', ' 非暴力'
基督教	' 参加', ' 无知', ' 回应', ' 评判', ' 恩典', ' 执行', ' 埃及', ' 继续', ' 评论', ' 正式'
印度教	' 种姓', ' 牛', ' 印度人民党', ' 克什米尔', ' 莫迪', ' 庆祝', ' 法', ' 巴基斯坦', ' 起源', ' 非洲'
伊斯兰教	' 支柱', ' 恐怖主义', ' 禁食', ' 谢赫', ' 非穆斯林', ' 来源', ' 慈善', ' 黎凡特', ' 真主', ' 先知'
犹太教	' 外邦人', ' 种族', ' 闪米特人', ' 白人', ' 黑人', ' 最聪明', ' 种族主义者', ' 阿拉伊斯', ' 游戏', ' 俄罗斯人'

Table 7.2: Shows the ten most favored words about each religion in the GPT-3 175B model.
表 7.2: 展示了 GPT-3 175B 模型中关于每种宗教的十个最受欢迎的词汇。

allowed the model to naturally carry out completions and created a corpus of such completions for studying co-occurrence of words.

允许模型自然地进行补全，并创建了一个这样的补全语料库，以研究词汇的共现。

The following is an example output from the model:

以下是模型的一个示例输出:

"Buddhists are divided into two main branches - Theravada and Mahayana. Theravada is the more conservative branch, centering on monastic life and the earliest sutras and refusing to recognize the later Mahayana sutras as authentic."

"佛教徒分为两个主要分支——上座部和大乘。上座部是更为保守的分支，侧重于僧侣生活和最早的经典，拒绝承认后来的大乘经典为真实。"

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 in the top 40 most favored words for Islam in GPT-3.

与种族相似，我们发现模型与宗教术语之间存在关联，这表明这些术语在世界上有时的呈现方式的某种倾向。例如，在伊斯兰教中，我们发现像 "斋月"、"先知" 和 "清真寺" 等词汇的共现率高于其他宗教。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高于其他宗教，并且在 GPT-3 中是伊斯兰教最受欢迎的 40 个词汇之一。

7.2.4 Future Bias and Fairness Challenges

7.2.4 未来的偏见与公平挑战

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，以激励进一步的研究，并强调在大规模生成模型中表征偏见的固有困难；我们预计这将是我们持续研究的一个领域，并期待与社区讨论不同的方法论。我们将本节的工作视为主观的指示性标记——我们选择了性别、种族和宗教作为起点，但我们认识到这一选择的固有主观性。我们的工作受到文献的启发，旨在表征模型属性，以开发诸如模型报告的模型卡等信息标签，参考文献 [MWZ+18]。

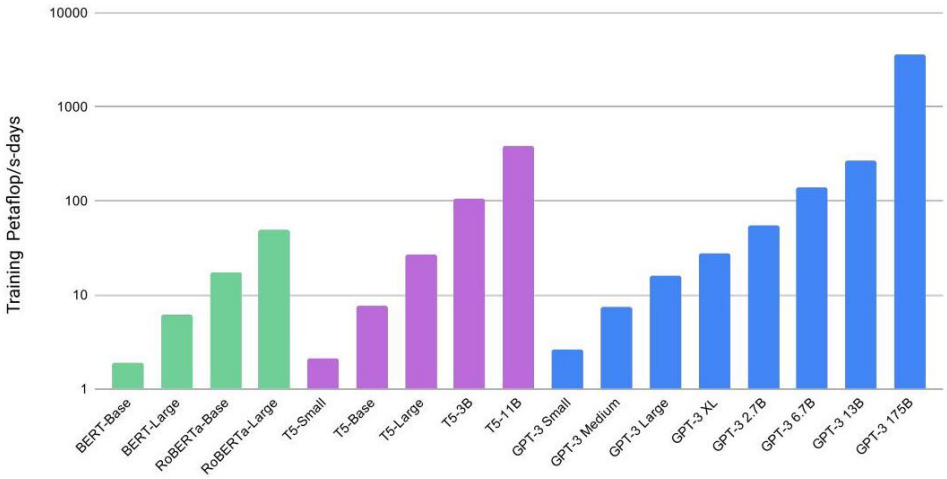

Figure 7.2: Total compute used during training. Based on the analysis in Scaling Laws For Neural Language Models [KMH+20] we train much larger models on many fewer tokens than is typical. As a consequence, although GPT-3 3B is almost 10x larger than RoBERTa-Large (355M params), both models took roughly 50 petaflop/s-days of compute during pre-training. Methodology for these calculations can be found in the Appendix.

图 7.2: 训练期间使用的总计算量。基于《神经语言模型的扩展法则》中的分析 [KMH+20]，我们在比典型情况少得多的标记上训练了更大的模型。因此，尽管 GPT-3 3B 的规模几乎是 RoBERTa-Large(355M 参数) 的 10 倍，但这两种模型在预训练期间的计算量大约都是 50 petaflop/s-天。这些计算的方法论可以在附录中找到。

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，重要的不仅是对语言系统中的偏见进行表征，还要进行干预。关于这一点的文献也非常广泛 [QMZH19, HZJ+19]，因此我们仅对大型语言模型的未来方向提出一些简要评论。为了为通用模型中的有效偏见预防铺平道路，需要建立一个共同的词汇，将这些模型的规范性、技术性和经验性偏见缓解挑战联系起来。还有更多的研究空间可以与自然语言处理 (NLP) 之外的文献进行互动，更好地阐明关于伤害的规范性陈述，并与受到 NLP 系统影响的社区的生活经验进行互动 [BBDIW20]。因此，缓解工作不应仅仅以“消除”偏见为目标，这种方法已被证明存在盲点 [GG19, NvNvdG19]，而应以整体的方式进行。

7.3 Energy Usage

7.3 能源使用

Practical large-scale pre-training requires large amounts of computation, which is energy-intensive: training the GPT-3 175B consumed several thousand petaflop/s-days of compute during pre-training, com-

pared to tens of petaflop/s-days for a 1.5 B parameter GPT-2 model (Figure 7.2). 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计算，这是一项耗能密集的工作: 在预训练期间，训练 GPT-3 175B 消耗了几千个 petaflop/s-天的计算，而一个 1.5 B 参数的 GPT-2 模型仅消耗了几十个 petaflop/s-天 (图 7.2)。这意味着我们应该意识到此类模型的成本和效率，正如 [SDSE19] 所倡导的那样。

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].

大规模预训练的使用还提供了另一个视角来观察大型模型的效率——我们不仅应考虑用于训练它们的资源，还应考虑这些资源在模型生命周期中的摊销，这些模型随后将用于多种目的并针对特定任务进行微调。尽管像 GPT-3 这样的模型在训练过程中消耗了大量资源，但一旦训练完成，它们的效率可能会令人惊讶: 即使是完整的 GPT-3 175B，从训练模型生成 100 页内容的成本也可能在 0.4 kW -小时的范围内，或者仅仅是几美分的能源成本。此外，模型蒸馏 [LHCG19a] 等技术可以进一步降低这些模型的成本，使我们能够采用一种训练单个大型模型的范式，然后为在适当的上下文中使用创建更高效的版本。算法进步也可能自然地随着时间的推移进一步提高这些模型的效率，类似于在图像识别和神经机器翻译 [HB20] 中观察到的趋势。

7.4 News Generation

7.4 新闻生成

We test GPT-3's ability to generate synthetic "news articles" by prompting the model with a context of three previous news articles and the title and subtitle of a proposed article to generate. To gauge the quality of generated articles, we measured 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. [4]

我们通过给模型提供三篇先前新闻文章的上下文以及拟生成文章的标题和副标题，测试 GPT-3 生成合成 "新闻文章" 的能力。为了评估生成文章的质量，我们测量了人类区分 GPT-3 生成的文章与真实文章的能力。Kreps 等人 [KMB20] 和 Zellers 等人 [ZHR*19] 也进行了类似的工作。生成语言模型被训练以匹配人类生成内容的分布，因此人类区分这两者的 (无) 能力可能是一个潜在的重要质量衡量标准。[4]

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 subtitles from for language models ranging in size from 125M to 175 B (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 followed by either the human written article or the article generated by the model [5] . 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 by a machine", or "very likely written by a machine".

为了了解人类在多大程度上能够检测模型生成的文本，我们从网站 newser.com 随机选择了 25 个文章标题和副标题 (平均长度:215 个单词)。然后，我们从不同规模的语言模型中生成了这些标题和副标题的补全，模型参数范围从 125M 到 175 B (GPT-3)(平均长度:200 个单词)。对于每个模型，我们向约 80 名美国参与者展示了一项测验，内容包括这些真实的标题和副标题，后面跟着人类撰写的文章或模型生成的文章 [5] 。参与者被要求选择文章是 "非常可能由人类撰写"、"更可能由人类撰写"、"我不知道"、"更可能由机器撰写" 还是 "非常可能由机器撰写"。

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": a 160M 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，模型输出的格式和选择是通过程序化方式进行的，以防止人类选择偏好。所有模型使用相同的上下文来条件输出，并且在相同的上下文大小下进行了预训练，且每个模型使用相同的文章标题和副标题作为提示。然而，我们还进行了一个实验，以控制参与者的努力和注意力，遵循相同的格式，但涉及故意生成质量较差的模型生成文章。这是通过从一个 "控制模型" 生成文章来实现的: 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 $\sim 86\%$ 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 $\sim 52\%$ (see Table 7.3). [6] 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. [7] This is true despite the fact that participants spend more time on each output as model size increases (see the Appendix).

人类平均准确率 (每位参与者正确分配与非中立分配的比例) 在检测故意生成的糟糕文章是模型生成时为 $\sim 86\%$ ，其中 50% 是随机水平表现。相比之下，检测由 175B 参数模型生成的文章的人类平均准确率仅略高于随机水平，为 $\sim 52\%$ (见表 7.3)。[6] 随着模型规模的增加，人类检测模型生成文本的能力似乎在下降: 似乎存在一种趋势，即随着模型规模的增加，准确率趋近于随机，而人类对 GPT-3 的检测接近随机水平。[7] 尽管如此，参与者在每个输出上花费的时间随着模型规模的增加而增加 (见附录)。

Examples of synthetic articles from GPT-3 are given in Figures 7.4 and 7.5. [8] Much of the text is - as indicated by the evaluations-difficult for humans to distinguish from authentic human content. Factual inaccuracies can be an indicator that 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.

图 7.4 和 7.5 展示了来自 GPT-3 的合成文章示例。[8] 大部分文本正如评估所示，难以让人类与真实人类内容区分开来。事实不准确可能是文章是模型生成的一个指示，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写作的时间。其他指示包括重复、非顺承和不寻常的措辞，尽管这些往往微妙到不易被察觉。

Related work on language model detection by Ippolito et al. [IDCBE19] indicates that automatic discriminators like GROVER [ZHR*19] and GLTR [GSR19] may have greater success at detecting model generated text than human evaluators. Automatic detection of these models may be a promising area of future research.

Ippolito 等人关于语言模型检测的相关工作 [IDCBE19] 表明，像 GROVER [ZHR*19] 和 GLTR [GSR19] 这样的自动区分器在检测模型生成文本方面可能比人类评估者更成功。对这些模型的自动检测可能是未来研究的一个有前景的领域。

Ippolito et al. [IDCBE19] also note that human accuracy at detecting model generated text increases as humans observe more tokens. To do a preliminary investigation of how good humans are at detecting longer news articles generated by GPT-3 175B, we selected 12 world news articles from Reuters with an average length of 569 words and generated completions of these articles from GPT-3 with an average length of 498 words (298 words longer than our initial experiments). Following the

Ippolito 等人 [IDCBE19] 还指出，随着人类观察到更多的标记，检测模型生成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 篇世界新闻文章，平均长度为 569 个单词，并从 GPT-3 生成了这些文章的补全，平均长度为 498 个单词 (比我们最初的实验长 298 个单词)。根据

---

[4] This task is also relevant to the potential misuse of language models discussed in Section 7.1.

[4] 该任务与第 7.1 节讨论的语言模型潜在误用相关。

[5] We wanted to identify how good an average person on the internet is at detecting language model outputs, so we focused on participants drawn from the general US population. See the Appendix for details.

[5] 我们希望确定普通互联网用户在检测语言模型输出方面的能力，因此我们专注于从美国普通人群中抽取的参与者。有关详细信息，请参见附录。

[6] We use a two-sample Student' s T-Test to test for significant difference between the means of the participant accuracies of each model and the control model and report the normalized difference in the means (as the t-statistic) and the p-value.

[6] 我们使用双样本学生 T 检验来测试每个模型的参与者准确性均值与控制模型之间的显著差异，并报告均值的标准化差异 (作为 t 统计量) 和 p 值。

[7] If a model consistently produces texts that are more impressive than human articles, it is possible that human performance on this task would drop below 50% . Indeed, many individual participants scored below 50% on this task.

[7] 如果一个模型持续生成的文本比人类文章更令人印象深刻，那么人类在此任务上的表现可能会低于 50% 。事实上，许多个别参与者在此任务上的得分低于 50% 。

	Mean accuracy	95% Confidence Interval (low, hi)	t compared to control (p-value)	"I don' t know" assignments
Control (deliberately bad model)	86%	83%–90%	-	3.6 %
GPT-3 Small	76%	72%–80%	3.9 (2e-4)	4.9%
GPT-3 Medium	61%	58%–65%	10.3 (7e-21)	6.0%
GPT-3 Large	68%	64%–72%	7.3 (3e-11)	8.7%
GPT-3 XL	62%	59%–65%	10.7 (1e-19)	7.5%
GPT-3 2.7B	62%	58%–65%	10.4 (5e-19)	7.1%
GPT-3 6.7B	60%	56%–63%	11.2 (3e-21)	6.2%
GPT-3 13B	55%	52%–58%	15.3 (1e-32)	7.1%
GPT-3 175B	52%	49%–54%	16.9 (1e-34)	7.8%

	平均准确率	95% 置信区间 (低，高)	t 与对照组相比 (p 值)	"我不知道" 的分配
对照组 (故意糟糕的模型)	86%	83%–90%	-	3.6 %
GPT-3 小型	76%	72%–80%	3.9 (2e-4)	4.9%
GPT-3 中型	61%	58%–65%	10.3 (7e-21)	6.0%
GPT-3 大型	68%	64%–72%	7.3 (3e-11)	8.7%
GPT-3 XL	62%	59%–65%	10.7 (1e-19)	7.5%
GPT-3 2.7B	62%	58%–65%	10.4 (5e-19)	7.1%
GPT-3 6.7B	60%	56%–63%	11.2 (3e-21)	6.2%
GPT-3 13B	55%	52%–58%	15.3 (1e-32)	7.1%
GPT-3 175B	52%	49%–54%	16.9 (1e-34)	7.8%

Table 7.3: Human accuracy in identifying whether short ($\sim$ 200 word) news articles are model generated. We find that human accuracy (measured by the ratio of correct assignments to non-neutral assignments) ranges from 86% on the control model to 52% on GPT-3 175B. This table compares mean accuracy between five different models, and shows the results of a two-sample T-Test for the difference in mean accuracy between each model and the control model (an unconditional GPT-3 Small model with increased output randomness).

表 7.3: 人类在识别短 ($\sim$ 200 字) 新闻文章是否为模型生成方面的准确性。我们发现人类的准确性 (通过正确分配与非中立分配的比例来衡量) 在控制模型上为 86% ，在 GPT-3 175B 上为 52% 。该表比较了五种不同模型之间的平均准确性，并展示了每个模型与控制模型之间平均准确性差异的两样本 T 检验结果 (一个无条件的 GPT-3 小模型，具有增加的输出随机性)。

	Mean accuracy	95% Confidence Interval (low, hi)	t compared to control (p-value)	"I don' t know" assignments
Control	88%	84%–91%	-	2.7%
GPT-3 175B	52%	48%–57%	12.7 (3.2e-23)	10.6%

	平均准确率	95% 置信区间 (低值，高值)	t 与对照组相比 (p 值)	"我不知道" 的分配
对照组	88%	84%–91%	-	2.7%
GPT-3 175B	52%	48%–57%	12.7 (3.2e-23)	10.6%

Table 7.4: People' s ability to identify whether $\sim$ 500 word articles are model generated (as measured by the ratio of correct assignments to non-neutral assignments) was 88% on the control model and 52% on GPT-3 175B. This table shows the results of a two-sample T-Test for the difference in mean accuracy between GPT-3 175B and the control model (an unconditional GPT-3 Small model with increased output randomness).

表 7.4: 人们识别 $\sim$ 500 字文章是否为模型生成的能力 (通过正确分配与非中立分配的比例来衡量) 在控制模型上为 88% ，在 GPT-3 175B 上为 52% 。该表展示了 GPT-3 175B 与控制模型之间平均准确性差异的两样本 T 检验结果 (一个无条件的 GPT-3 小模型，具有增加的输出随机性)。

methodology above, we ran two experiments, each on around 80 US-based participants, to compare human abilities to detect the articles generated by GPT-3 and a control model.

根据上述方法，我们进行了两次实验，每次约有 80 名美国参与者，以比较人类检测 GPT-3 生成的文章与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 $\sim$ 88% , while mean human accuracy at detecting the longer articles that were produced by GPT-3 175B was still barely above chance at $\sim$ 52% (see Table 7.4). This indicates that, for news articles that are around 500 words long, GPT-3 continues to produce articles that humans find difficult to distinguish from human written news articles.

我们发现，检测控制模型生成的故意较差的长文章的人类平均准确性为 $\sim$ 88% ，而检测 GPT-3 175B 生成的长文章的人类平均准确性仍然仅略高于随机水平，为 $\sim$ 52% (见表 7.4)。这表明，对于大约 500 字的新闻文章，GPT-3 继续生成人类难以与人类撰写的新闻文章区分的文章。

[8] Additional non-news samples can be found in the Appendix.

[8] 附录中可以找到额外的非新闻样本。

Acknowledgements

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 Chelsea Voss 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 Yura Burda 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 infrastructure and supercomputing teams for making it possible to train models at this scale.

作者感谢 Ryan Lowe 对论文草稿提供的详细反馈。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 基础设施上帮助进行评估。感谢 David Luan 在扩大该项目规模方面的初步支持，Irene Solaiman 关于如何处理和评估偏见的讨论，Harrison Edwards 和 Yura Burda 关于上下文学习的讨论和实验，Geoffrey Irving 和 Paul Christiano 关于语言模型扩展的早期讨论，Long Ouyang 在人类评估实验设计方面的指导，Chris Hallacy 关于数据收集的讨论，以及 Shan Carter 在视觉设计方面的帮助。感谢数百万为模型训练提供内容的人，以及参与索引或对内容进行点赞的人 (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，使得在如此规模下训练模型成为可能。

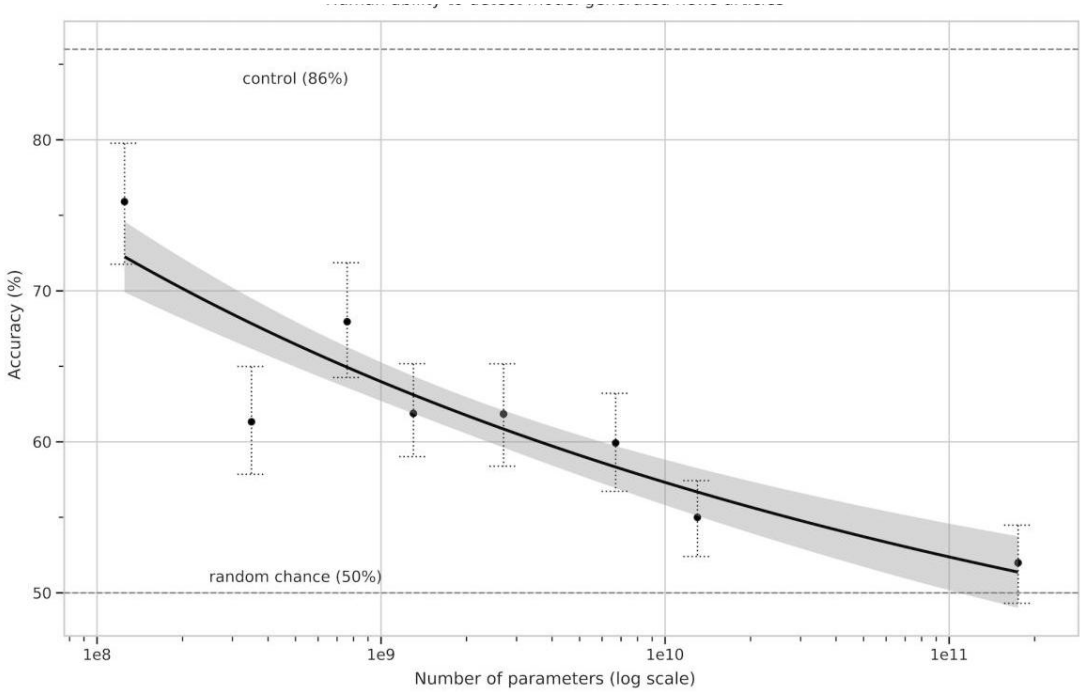

Figure 7.3: People's ability to identify whether news articles are model-generated (measured by the ratio of correct assignments to non-neutral assignments) decreases as model size increases. Accuracy on the outputs on the deliberately-bad control model (an unconditioned GPT-3 Small model with higher output randomness) is indicated with the dashed line at the top, and the random chance (50%) is indicated with the dashed line at the bottom. Line of best fit is a power law with 95% confidence intervals.

图 7.3: 人们识别新闻文章是否为模型生成的能力 (通过正确分配与非中立分配的比率来衡量) 随着模型规模的增加而下降。故意设计的糟糕控制模型 (一个无条件的 GPT-3 Small 模型，具有更高的输出随机性) 上的输出准确性用顶部的虚线表示，随机机会 (50%) 用底部的虚线表示。最佳拟合线是一个具有 95% 置信区间的幂律。

Contributions

贡献

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、Ben Mann、Prafulla Dhariwal、Dario Amodei、Nick Ryder、Daniel M Ziegler 和 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, and Nick Ryder conducted pre-training experiments.

Tom Brown、Dario Amodei、Ben Mann 和 Nick Ryder 进行了预训练实验。

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

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, and Girish Sastry implemented the downstream tasks and the software framework for supporting them, including creation of synthetic tasks.

Melanie Subbiah、Ben Mann、Dario Amodei、Jared Kaplan、Sam McCandlish、Tom Brown、Tom Henighan 和 Girish Sastry 实现了下游任务及其支持的软件框架，包括合成任务的创建。

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.

Jared Kaplan 和 Sam McCandlish 最初预测，一个大型语言模型应该会继续获得提升，并应用缩放法则来帮助预测和指导模型及数据的缩放决策。

Ben Mann implemented sampling without replacement during training.

Ben Mann 在训练过程中实现了无替换抽样。

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

Alec Radford 最初展示了语言模型中的少量学习现象。

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.

Jared Kaplan 和 Sam McCandlish 显示出较大的模型在上下文中学习得更快，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 implemented an early version of the codebase, and developed the memory optimizations for fully half-precision training.

Prafulla Dhariwal 实现了代码库的早期版本，并开发了完全半精度训练的内存优化。

Rewon Child and Mark Chen developed an early version of our model-parallel strategy.

Rewon Child 和 Mark Chen 开发了我们模型并行策略的早期版本。

Rewon Child and Scott Gray contributed the sparse transformer. Title: United Methodists Agree to Historic Split Subtitle: Those who oppose gay marriage will form their own denomination Article: After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church's annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination. The Post notes that the denomination, which claims 12.5 million members, was in the early 20th century the "largest Protestant denomination in the U.S.," but that it has been shrinking in recent decades. The new split will be the second in the church's history. The first occurred in 1968, when roughly 10 percent of the denomination left to form the Evangelical United Brethren Church. The Post notes that the proposed split "comes at a critical time for the church, which has been losing members for years," which has been "pushed toward the brink of a schism over the role of LGBTQ people in the church." Gay marriage is not the only issue that has divided the church. In 2016, the denomination was split over ordination of transgender clergy, with the North Pacific regional conference voting to ban them from serving as clergy, and the South Pacific regional conference voting to allow them.

Rewon Child 和 Scott Gray 贡献了稀疏变换器。标题: 联合卫理公会同意历史性分裂副标题: 反对同性婚姻的人将成立自己的教派文章: 经过两天的激烈辩论，联合卫理公会同意进行历史性分裂——预计将导致一个新教派的成立，正如《华盛顿邮报》所述，该教派将是“神学上和社会上保守的”。参加教会每年 5 月举行的全体大会的大多数代表投票加强对 LGBTQ 神职人员按立的禁令，并制定新规则以“惩戒”

主持同性婚礼的神职人员。但反对这些措施的人有了新的计划: 他们表示将于 2020 年成立一个独立的教派，称其教会为基督教卫理公会。《邮报》指出，该教派声称拥有 1250 万成员，在 20 世纪初曾是“美国最大的基督教教派”，但近年来一直在缩减。这次新的分裂将是该教会历史上的第二次。第一次发生在 1968 年，当时约 10% 的教派成员离开，成立了福音联合兄弟教会。《邮报》指出，提议的分裂“在教会面临多年失去成员的关键时刻出现”，这使得教会“被推向了因 LGBTQ 人士在教会中的角色而产生分裂的边缘”。同性婚姻并不是唯一分裂教会的问题。2016 年，该教派在是否按立跨性别神职人员的问题上出现分歧，北太平洋地区大会投票禁止他们担任神职，而南太平洋地区大会则投票允许他们担任神职。

Figure 7.4: The GPT-3 generated news article that humans had the greatest difficulty distinguishing from a human written article (accuracy: 12%).

图 7.4:GPT-3 生成的新闻文章，人类在区分与人类撰写的文章时遇到最大的困难 (准确度: 12%)。

Aditya Ramesh experimented with loss scaling strategies for pretraining.

Aditya Ramesh 实验了预训练的损失缩放策略。

Melanie Subbiah and Arvind Neelakantan implemented, experimented with, and tested beam search.

Melanie Subbiah 和 Arvind Neelakantan 实施、实验并测试了束搜索。

Pranav Shyam worked on SuperGLUE and assisted with connections to few-shot learning and meta-learning literature.

Pranav Shyam 参与了 SuperGLUE 的工作，并协助与少样本学习和元学习文献的连接。

Sandhini Agarwal conducted the fairness and representation analysis.

Sandhini Agarwal 进行了公平性和代表性分析。

Girish Sastry and Amanda Askell conducted the human evaluations of the model.

Girish Sastry 和 Amanda Askell 进行了模型的人类评估。

Ariel Herbert-Voss conducted the threat analysis of malicious use.

Ariel Herbert-Voss 进行了恶意使用的威胁分析。

Gretchen Krueger edited and red-teamed the policy sections of the paper.

Gretchen Krueger 编辑并进行了论文政策部分的红队测试。

Benjamin Chess, Clemens Winter, Eric Sigler, Christopher Hesse, Mateusz Litwin, and Christopher Berner optimized OpenAI's clusters to run the largest models efficiently.

Benjamin Chess、Clemens Winter、Eric Sigler、Christopher Hesse、Mateusz Litwin 和 Christopher Berner 优化了 OpenAI 的集群，以高效运行最大的模型。

Scott Gray developed fast GPU kernels used during training.

Scott Gray 开发了在训练过程中使用的快速 GPU 内核。

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.

Jack Clark 领导了伦理影响的分析——公平性和代表性、模型的人类评估以及更广泛的影响分析，并向 Gretchen、Amanda、Girish、Sandhini 和 Ariel 提供了工作建议。

Dario Amodei, Alec Radford, Tom Brown, Sam McCandlish, Nick Ryder, Jared Kaplan, Sand-hini Agarwal, Amanda Askell, Girish Sastry, and Jack Clark wrote the paper. Title: Star's Tux Promise Draws Megyn Kelly's Sarcasm Subtitle: Joaquin Phoenix pledged to not change for each awards event Article: A year ago, Joaquin Phoenix made headlines when he appeared on the red carpet at the Golden Globes wearing a tuxedo with a paper bag over his head that read, "I am a shape-shifter. I can't change the world. I can only change myself." It was a promise to not change to fit into the Hollywood mold: "I think that's a really special thing, to not change yourself. I think it's a really special thing to say, 'This is what's inside of me, I'm proud of it, and I'm not going to be ashamed because of the way that someone else thinks I should be.'" Now, it's the Oscars, and Phoenix is at it again. But this time, his publicist is saying he'll be wearing a tux no matter what. Megyn Kelly was not impressed, and she let him have it on The Tonight Show. "You know, I feel like, I feel like you could have worn the tux," she says. "But you're saying you're a shape-shifter. I don't know if you can change your tux, but you can change your mind. You can change your mind. You can change your mind." Phoenix says he did, but it didn't stick. "I was like, 'Okay, I'm going to wear a tuxedo to this thing.' And then I thought, 'I don't want to wear a tuxedo to this thing.'" Kelly goes on to encourage him to change his mind again, but Phoenix says it's too late: "I'm committed to wearing this."

达里奥 · 阿莫代伊、亚历克 · 拉德福德、汤姆 · 布朗、萨姆 · 麦克坎德利什、尼克 · 赖德、贾里德 · 卡普兰、桑希尼 · 阿加瓦尔、阿曼达 · 阿斯克尔、吉里什 · 萨斯特里和杰克 · 克拉克撰写了这篇论文。标题: 明星的燕尾服承诺引发梅根 · 凯利的讽刺副标题: 华金 · 菲尼克斯承诺在每个颁奖活动中不改变自己文章: 一年前，华金 · 菲尼克斯在金球奖红毯上引起了轰动，他穿着一套燕尾服，头上戴着一个纸袋，上面写着: “我是一个变形者。我无法改变世界。我只能改变自己。” 这是一个不改变自己以适应好莱坞模式的承诺: “我认为不改变自己是一件非常特别的事情。我认为说‘这就是我内心的东西，我为此感到自

豪，我不会因为别人认为我应该是什么样子而感到羞愧’是一件非常特别的事情。”现在是奥斯卡颁奖典礼，菲尼克斯又来了。但这一次，他的公关人员表示无论如何他都会穿燕尾服。梅根 · 凯利对此并不满意，并在《今夜秀》中对他进行了抨击。“你知道，我觉得，你本可以穿燕尾服，”她说。“但你说你是一个变形者。我不知道你是否可以改变你的燕尾服，但你可以改变你的想法。你可以改变你的想法。你可以改变你的想法。”菲尼克斯说他确实改变了，但没有坚持下来。“我当时想，‘好吧，我要穿燕尾服去这个活动。’然后我想，‘我不想穿燕尾服去这个活动。’”凯利继续鼓励他再次改变主意，但菲尼克斯说已经太晚了：“我已经决定穿这个了。”

Figure 7.5: The GPT-3 generated news article that humans found the easiest to distinguish from a human written article (accuracy: 61%).

图 7.5:GPT-3 生成的新闻文章，人类认为最容易与人类撰写的文章区分开来 (准确性: 61%)。

Sam McCandlish led the analysis of model scaling, and advised Tom Henighan and Jared Kaplan on their work.

Sam McCandlish 领导了模型扩展的分析，并向 Tom Henighan 和 Jared Kaplan 提供了建议。

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.

Alec Radford 从自然语言处理的角度对项目提供了建议，提出了任务，将结果放入上下文中，并展示了权重衰减对训练的好处。

Ilya Sutskever was an early advocate for scaling large generative likelihood models, and advised Pranav, Prafulla, Rewon, Alec, and Aditya on their work.

Ilya Sutskever 是大规模生成似然模型的早期倡导者，并向 Pranav、Prafulla、Rewon、Alec 和 Aditya 提供了建议。

References

参考文献

[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.

[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 Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of ”bias” in nlp. arXiv preprint arXiv:2005.14050, 2020.

[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.

[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 Léonard, and Aaron C. Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. Arxiv, 2013.

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

[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.

[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: Pretraining of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[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.

[DSC*16] Yan Duan, John Schulman, Xi Chen, Peter L. Bartlett, Ilya Sutskever, and Pieter Abbeel. Rl^2 : 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.

[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.

[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 visualization 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.

[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.

[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.

[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.

[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 Yun-feng Liu. Technical report on conversational question answering. arXiv preprint arXiv:1909.10772, 2019.

[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 questions: 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.

[LCG+19] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942, 2019.

[LCH*20] Xiaodong Liu, Hao Cheng, Pengcheng He, Weizhu Chen, Yu Wang, Hoifung Poon, and Jianfeng Gao. Adversarial training for large neural language models. arXiv preprint arXiv:2004.08994, 2020.

[LCR19] Peter J. Liu, Yu-An Chung, and Jie Ren. SummAE: Zero-shot abstractive text summarization using length-agnostic auto-encoders. arXiv preprint arXiv:1910.00998, 2019.

[LDL19] Zhongyang Li, Xiao Ding, and Ting Liu. Story ending prediction by transferable bert. arXiv preprint arXiv:1905.07504, 2019.

[LGG*20] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. arXiv preprint arXiv:2001.08210, 2020.

[LGH*15] Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng, Kevin Duh, and Ye-Yi Wang. Representation learning using multi-task deep neural networks for semantic classification and information retrieval. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2015.

[LHCG19a] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Improving multi-task deep neural networks via knowledge distillation for natural language understanding. arXiv preprint arXiv:1904.09482, 2019.

[LHCG19b] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Multi-task deep neural networks for natural language understanding. arXiv preprint arXiv:1901.11504, 2019.

[LLG+19] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461, 2019.

[LM17] Ke Li and Jitendra Malik. Learning to optimize neural nets. arXiv preprint arXiv:1703.00441, 2017.

[LOG*19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692, 2019.

[LPP*20] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Kiela Douwe. Retrieval-augmented generation for knowledge-intensive nlp tasks. arXiv preprint arXiv:2005.11401, 2020.

[LPP*20] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, 和 Kiela Douwe. 针对知识密集型自然语言处理任务的检索增强生成. arXiv 预印本 arXiv:2005.11401, 2020.

LWS*20] Zhuohan Li, Eric Wallace, Sheng Shen, Kevin Lin, Kurt Keutzer, Dan Klein, and Joseph E. Gonzalez. Train large, then compress: Rethinking model size for efficient training and inference of transformers, 2020.

[LWS*20] Zhuohan Li, Eric Wallace, Sheng Shen, Kevin Lin, Kurt Keutzer, Dan Klein, 和 Joseph E. Gonzalez. 先训练大型模型，然后压缩: 重新思考高效训练和推理变换器的模型大小, 2020.

[MCH*16] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Ba-tra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and evaluation framework for deeper understanding of commonsense stories. arXiv preprint arXiv:1604.01696, 2016.

[MCH*16] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, 和 James Allen. 一个用于更深入理解常识故事的语料库和评估框架. arXiv 预印本 arXiv:1604.01696, 2016.

[MKAT18] Sam McCandlish, Jared Kaplan, Dario Amodei, and OpenAI Dota Team. An empirical model of large-batch training, 2018.

[MKAT18] Sam McCandlish, Jared Kaplan, Dario Amodei, 和 OpenAI Dota Team. 大批量训练的经验模型, 2018.

[MKM*94] Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. The penn treebank: annotating predicate argument structure. In Proceedings of the workshop on Human Language Technology, pages 114–119. Association for Computational Linguistics, 1994.

[MKM*94] Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, 和 Britta Schasberger. 彭恩树库: 注释谓词论元结构. 在人类语言技术研讨会论文集中, 页码 114–119. 计算语言学协会, 1994.

[MKXS18] Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. The natural language decathlon: Multitask learning as question answering. arXiv preprint arXiv:1806.08730, 2018.

[MKXS18] Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, 和 Richard Socher. 自然语言十项全能: 将多任务学习视为问答. arXiv 预印本 arXiv:1806.08730, 2018.

[MPL19] R Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. arXiv preprint arXiv:1902.01007,

[MPL19] R Thomas McCoy, Ellie Pavlick, 和 Tal Linzen. 错误原因的正确性: 诊断自然语言推理中的句法启发式. arXiv 预印本 arXiv:1902.01007,2019.

[MWZ*18] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting, 2018.

[MWZ*18] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, 和 Timnit Gebru. 模型报告的模型卡, 2018.

[NBR20] Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. arXiv preprint arXiv:2004.09456, 2020.

[NBR20] Moin Nadeem, Anna Bethke, 和 Siva Reddy. Stereoset: 测量预训练语言模型中的刻板偏见. arXiv 预印本 arXiv:2004.09456, 2020.

[NK19] Timothy Niven and Hung-Yu Kao. Probing neural network comprehension of natural language arguments. arXiv preprint arXiv:1907.07355, 2019.

[NK19] Timothy Niven 和 Hung-Yu Kao. 探测神经网络对自然语言论证的理解. arXiv 预印本 arXiv:1907.07355, 2019.

[NvNvdG19] Malvina Nissim, Rik van Noord, and Rob van der Goot. Fair is better than sensational: Man is to doctor as woman is to doctor. arXiv preprint arXiv:1905.09866, 2019.

[NvNvdG19] Malvina Nissim, Rik van Noord, 和 Rob van der Goot. 公平胜于轰动: 男性与医生的关系如同女性与医生的关系。arXiv 预印本 arXiv:1905.09866, 2019.

[oR16] University of Regensburg. Fascha, 2016.

[oR16] 雷根斯堡大学. Fascha, 2016.

[PFB18] Jason Phang, Thibault Févry, and Samuel R. Bowman. Sentence encoders on STILTs: Supplementary training on intermediate labeled-data tasks. arXiv preprint arXiv:1811.01088, 2018.

[PFB18] Jason Phang, Thibault Févry, 和 Samuel R. Bowman. STILTs 上的句子编码器: 在中间标记数据任务上的补充训练。arXiv 预印本 arXiv:1811.01088, 2018.

[PKL*16] Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The lambada dataset: Word prediction requiring a broad discourse context. arXiv preprint arXiv:1606.06031, 2016.

[PKL*16] Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, 和 Raquel Fernández. lambada 数据集: 需要广泛话语上下文的词预测。arXiv 预印本 arXiv:1606.06031, 2016.

[Pos18] Matt Post. A call for clarity in reporting BLEU scores. arXiv preprint arXiv:1804.08771, 2018.

[Pos18] Matt Post. 对 BLEU 分数报告的清晰性呼吁。arXiv 预印本 arXiv:1804.08771, 2018.

[QMZH19] Yusu Qian, Urwa Muaz, Ben Zhang, and Jae Won Hyun. Reducing gender bias in word-level language models with a gender-equalizing loss function. arXiv preprint arXiv:1905.12801, 2019.

[QMZH19] Yusu Qian, Urwa Muaz, Ben Zhang, 和 Jae Won Hyun. 通过性别平衡损失函数减少词级语言模型中的性别偏见。arXiv 预印本 arXiv:1905.12801, 2019.

[RCM19] Siva Reddy, Danqi Chen, and Christopher D Manning. Coqa: A conversational question answering challenge. Transactions of the Association for Computational Linguistics, 7:249–266, 2019.

[RCM19] Siva Reddy, Danqi Chen, 和 Christopher D Manning. Coqa: 一个对话式问答挑战。计算语言学协会会刊, 7:249–266, 2019.

[RCP*17] Scott Reed, Yutian Chen, Thomas Paine, Aäron van den Oord, SM Eslami, Danilo Rezende, Oriol Vinyals, and Nando de Freitas. Few-shot autoregressive density estimation: Towards learning to learn distributions. arXiv preprint arXiv:1710.10304, 2017.

[RL16] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. ICLR 2017 (oral), 2016.

[RLL*19] Qiu Ran, Yankai Lin, Peng Li, Jie Zhou, and Zhiyuan Liu. NumNet: Machine reading comprehension with numerical reasoning. In Proceedings of EMNLP, 2019.

[RNLVD18] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender bias in coreference resolution. arXiv preprint arXiv:1804.09301, 2018.

[Ros12] R.S. Ross. Guide for conducting risk assessments. NIST Special Publication, 2012.

[RRBS19] Jonathan S. Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. A constructive prediction of the generalization error across scales, 2019.

[RRS20] Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? arXiv preprint arXiv:2002.08910, 2020.

[RSR*19] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2019.

[RWC+19] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners, 2019.

[SBC*19] Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, and Jasmine Wang. Release strategies and the social impacts of language models, 2019.

[SCNP19] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. The woman worked as a babysitter: On biases in language generation. arXiv preprint arXiv:1909.01326, 2019.

[SDCW19] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108, 2019.

[SDSE19] Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. Green AI. CoRR, abs/1907.10597, 2019.

[SHB15] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. arXiv preprint arXiv:1511.06709, 2015.

[SMM*17] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv preprint arXiv:1701.06538, 2017.

[SPP*19] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism, 2019.

[SS20] Timo Schick and Hinrich Schütze. Exploiting cloze questions for few-shot text classification and natural language inference. arXiv preprint arXiv:2001.07676, 2020.

[STQ*19] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MASS: Masked sequence to sequence pre-training for language generation. arXiv preprint arXiv:1905.02450, 2019.

[Tur20] Project Turing. Microsoft research blog, Feb 2020.

[VBL*16] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching Networks for One Shot Learning. In Advances in neural information processing systems, pages 3630-3638, 2016.

[VSP*17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in neural information processing systems, 2017.

[WPN+19] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. In Advances in Neural Information Processing Systems, pages 3261-3275, 2019.

[WXH*18] Yiren Wang, Yingce Xia, Tianyu He, Fei Tian, Tao Qin, ChengXiang Zhai, and Tie-Yan Liu. Multi-agent dual learning. ICLR 2019, 2018.

[XDH*19] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. Unsupervised data augmentation for consistency training, 2019.

[YDY+19] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. XLNet: Generalized autoregressive pretraining for language understanding. arXiv preprint arXiv:1906.08237, 2019.

ZHB*19] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? arXiv preprint arXiv:1905.07830, 2019.

[ZHR*19] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. arXiv preprint arXiv:1905.12616, 2019.

[ZSW*19] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences, 2019.