

Language Models are Few-Shot Learners

语言模型是少样本学习

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

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*
 Jared Kaplan† 普拉富拉·达里瓦尔 Arvind Neelakantan 普拉纳夫·谢姆 Girish Sastry
 Amanda Askell Sandhini Agarwal Ariel Herbert-Voss 格雷琴·克鲁格 汤姆·亨尼汉
 里温·查尔德 阿迪亚·拉梅什 丹尼尔·M·齐格勒 杰弗里·吴 克莱门斯·温特
 克里斯托弗·赫塞 马克·陈 埃里克·西格尔 马特乌什·利特温 斯科特·格雷
 本杰明·切斯 杰克·克拉克 克里斯托弗·伯纳
 山姆·麦克坎迪什 亚历克·拉德福德 伊利亚·苏茨凯弗 达里奥·阿莫迪

OpenAI

OpenAI

Abstract

摘要

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

近期研究表明，通过在大型文本语料库上进行预训练，然后针对特定任务进行微调，许多自然语言处理（NLP）任务和基准测试都取得了显著提升。虽然这种方法在架构上通常是任务无关的，但仍然需要数千或数万个特定任务的微调数据集。相比之下，人类通常只需要几个示例或简单的指令就能完成新的语言任务——而当前的NLP系统在这方面仍然面临很大挑战。在这里，我们展示了扩大语言模型规模如何显著提高任务无关的少样本性能，有时甚至可以达到先前最先进的微调方法的竞争力。具体来说，我们训练了GPT-3，一个拥有1750亿参数的自回归语言模型，比任何之前的非稀疏语言模型多10倍，并在少样本设置中测试其性能。对于所有任务，GPT-3在不进行任何梯度更新或微调的情况下应用，任务和少样本演示完全通过文本交互指定。GPT-3在许多NLP数据集上表现出色，包括翻译、问答和完形填空任务，以及一些需要即时推理或领域适应的任务，例如解乱词、在句子中使用新词或执行3位数字运算。同时，我们也发现GPT-3的少样本学习在某些数据集上仍然存在困难，以及一些数据集存在与在大型网络语料库上训练相关的方法论问题。最后，我们发现GPT-3可以生成新闻文章的样本，人类评估者难以将其与人类撰写的文章区分开来。我们讨论了这一发现的更广泛的社会影响以及GPT-3的总体影响。

*Equal contribution

†Johns Hopkins University, OpenAI

*同等贡献

†约翰霍普金斯大学，OpenAI

Contents

1 Introduction	3
2 Approach	6
2.1 Model and Architectures	8
2.2 Training Dataset	8
2.3 Training Process	9
2.4 Evaluation	10
3 Results	10
3.1 Language Modeling, Cloze, and Completion Tasks	11
3.2 Closed Book Question Answering	13
3.3 Translation	14
3.4 Winograd-Style Tasks	16
3.5 Common Sense Reasoning	17
3.6 Reading Comprehension	18
3.7 SuperGLUE	18
3.8 NLI	20
3.9 Synthetic and Qualitative Tasks	21
4 Measuring and Preventing Memorization Of Benchmarks	29
5 Limitations	33
6 Broader Impacts	34
6.1 Misuse of Language Models	35
6.2 Fairness, Bias, and Representation	36
6.3 Energy Usage	39
7 Related Work	39
8 Conclusion	40
A Details of Common Crawl Filtering	43
B Details of Model Training	43
C Details of Test Set Contamination Studies	43
D Total Compute Used to Train Language Models	46
E Human Quality Assessment of Synthetic News Articles	46
F Additional Samples from GPT-3	48
G Details of Task Phrasing and Specifications	50
H Results on All Tasks for All Model Sizes	63

目录

1 引言	3
2 方法	6
2.1 模型和架构	8
2.2 训练数据集	8
2.3 训练过程	9
2.4 评估	10
3 结果	10
3.1 语言建模、完形填空和补全任务	11
3.2 闭卷问答	13
3.3 翻译	14
3.4 Winograd风格任务	16
3.5 常识推理	17
3.6 阅读理解	18
3.7 SuperGLUE	18
3.8 NLI	20
3.9 合成和定性任务	21
4 测量和防止基准测试记忆	29
5 限制	33
6 更广泛的影响	34
6.1 语言模型的滥用	35
6.2 公平性、偏见和代表性	36
6.3 能源使用	39
7 相关工作	39
8 结论	40
A 常见爬虫过滤器细节	43
B 模型训练细节	43
C 测试集污染研究细节	43
D 训练语言模型所需总计算资源	46
E 合成新闻文章的人类质量评估	46
F 来自GPT-3的额外样本	48
G 任务措辞和规范细节	50
H 所有模型尺寸在所有任务上的结果	63

1 Introduction

Recent years have featured a trend towards pre-trained language representations in NLP systems, applied in increasingly flexible and task-agnostic ways for downstream transfer. First, single-layer representations were learned using word vectors [MCCD13, PSM14] and fed to task-specific architectures, then RNNs with multiple layers of representations and contextual state were used to form stronger representations [DL15, MBXS17, PNZY18] (though still applied to task-specific architectures), and more recently pre-trained recurrent or transformer language models [VSP+17] have been directly fine-tuned, entirely removing the need for task-specific architectures [RNSS18, DCLT18, HR18].

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

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

Second, the potential to exploit spurious correlations in training data fundamentally grows with the expressiveness of the model and the narrowness of the training distribution. This can create problems for the pre-training plus fine-tuning paradigm, where models are designed to be large to absorb information during pre-training, but are then fine-tuned on very narrow task distributions. For instance [HLW+20] observe that larger models do not necessarily generalize better out-of-distribution. There is evidence that suggests that the generalization achieved under this paradigm can be poor because the model is overly specific to the training distribution and does not generalize well outside it [YdC+19, MPL19]. Thus, the performance of fine-tuned models on specific benchmarks, even when it is nominally at human-level, may exaggerate actual performance on the underlying task [GSL+18, NK19].

Third, humans do not require large supervised datasets to learn most language tasks – a brief directive in natural language (e.g. “please tell me if this sentence describes something happy or something sad”) or at most a tiny number of demonstrations (e.g. “here are two examples of people acting brave; please give a third example of bravery”) is often

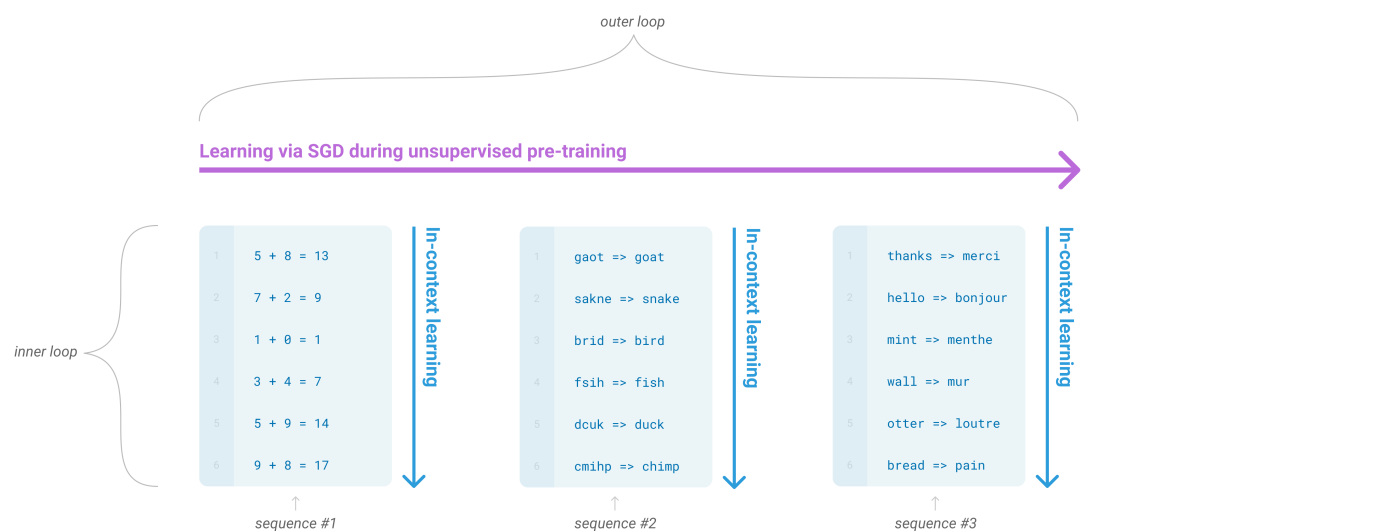


Figure 1.1: Language model meta-learning. During unsupervised pre-training, a language model develops a broad set of skills and pattern recognition abilities. It then uses these abilities at inference time to rapidly adapt to or recognize the desired task. We use the term “in-context learning” to describe the inner loop of this process, which occurs within the forward-pass upon each sequence. The sequences in this diagram are not intended to be representative of the data a model would see during pre-training, but are intended to show that there are sometimes repeated sub-tasks embedded within a single sequence.

1 简介

近年来，自然语言处理系统中的预训练语言表示呈现出一种趋势，以越来越灵活和任务无关的方式应用于下游迁移。首先，使用词向量学习单层表示 [MCCD13, PSM14]，然后将其输入到特定任务的架构中，接着使用具有多层表示和上下文状态的RNN形成更强的表示 [DL15, MBXS17, PNZY18] (尽管仍然应用于特定任务的架构)，以及最近预训练的循环或Transformer语言模型 [VSP+17]被直接微调，完全消除了对特定任务架构的需求 [RNSS18, DCLT18, HR18]。

这一最新范式在许多具有挑战性的自然语言处理任务（如阅读理解、问答、文本蕴涵等）上取得了实质性进展，并且基于新的架构和算法持续发展 [RSR+19, LOG+19, YDY+19, LCG+19]。然而，这种方法的一个主要局限性在于，虽然架构是任务无关的，但仍然需要特定任务的数集和特定任务的微调：要在期望的任务上实现强大性能，通常需要在数千到数十万个特定于该任务的示例数据集上进行微调。消除这一局限性是值得期望的，原因有数个。

首先，从实际角度来看，每个新任务都需要大量标记示例的需求限制了语言模型的应用。存在非常广泛的可能的有用语言任务，涵盖从纠正语法、生成抽象概念的示例到评论短篇小说等一切。对于许多这些任务，收集一个大的监督训练数据集很困难，尤其是在必须针对每个新任务重复该过程时。

其次，利用训练数据中的虚假关联的潜力，随着模型的表达能力和训练分布的狭窄程度增加而根本性增长。这可能会给预训练加微调的范式带来问题，在这种范式中，模型被设计得很大，以便在预训练期间吸收信息，但随后在非常狭窄的任务分布上进行微调。例如 [HLW+20] 观察到，更大的模型并不一定泛化得更好。有证据表明，在这种范式下实现的泛化可能很差，因为模型对训练分布过于特定，并且在其外部泛化不佳 [YdC+19, MPL19]。因此，微调模型在特定基准测试上的性能，即使名义上达到人类水平，也可能夸大其在底层任务上的实际性能 [GSL+18, NK19]。

第三，人类学习大多数语言任务并不需要大规模的监督数据集——一个自然语言的简短指令（例如“请告诉我，这个句子描述的是快乐还是悲伤”）或最多只需少量示范（例如“这里有两个人表现勇敢的例子；请给出第三个勇敢的例子”）就足够让人类能够至少达到合理程度的胜任新任务。除了指出我们当前自然语言处理技术中的一个概念性局限外，这种适应性还有实际优势——它允许人类能够无缝地混合或切换多种任务和技能，例如在长时间的对话中进行加法运算。为了广泛地发挥作用，我们希望有一天我们的自然语言处理系统能够拥有同样的灵活性和通用性。

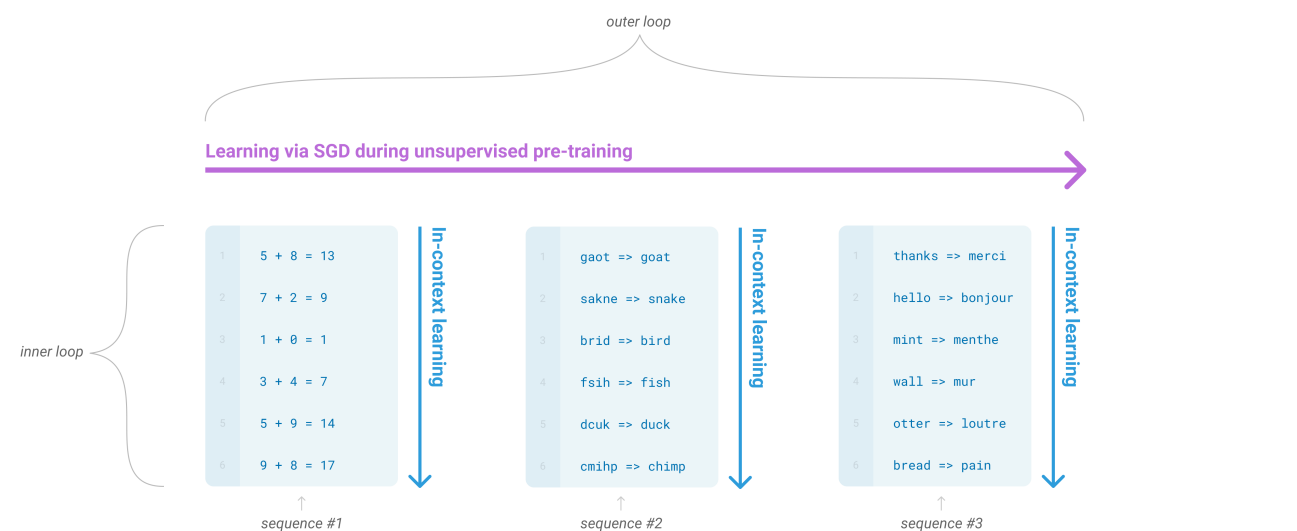


图1.1: 语言模型元学习。 在无监督预训练期间，语言模型发展出一套广泛的能力和模式识别能力。然后它利用这些能力在推理时间快速适应或识别所需的任务。我们使用“上下文学习”一词来描述此过程的内部循环，该循环在每个序列的正向传递中发生。此图表中的序列并非旨在代表模型在预训练期间会看到的数据，而是旨在表明有时单个序列中嵌入了重复的子任务。

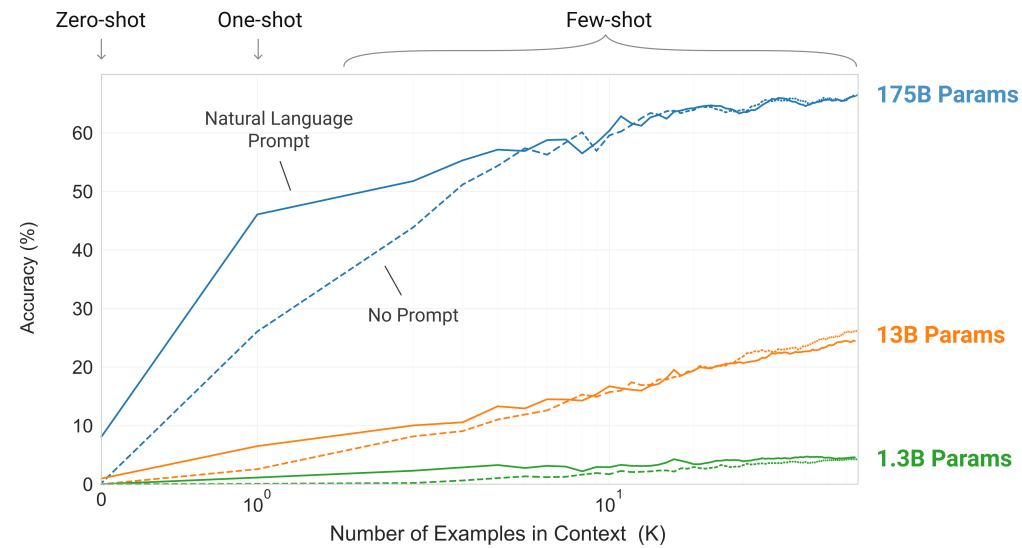


Figure 1.2: Larger models make increasingly efficient use of in-context information. We show in-context learning performance on a simple task requiring the model to remove random symbols from a word, both with and without a natural language task description (see Sec. 3.9.2). The steeper “in-context learning curves” for large models demonstrate improved ability to learn a task from contextual information. We see qualitatively similar behavior across a wide range of tasks.

sufficient to enable a human to perform a new task to at least a reasonable degree of competence. Aside from pointing to a conceptual limitation in our current NLP techniques, this adaptability has practical advantages – it allows humans to seamlessly mix together or switch between many tasks and skills, for example performing addition during a lengthy dialogue. To be broadly useful, we would someday like our NLP systems to have this same fluidity and generality.

One potential route towards addressing these issues is meta-learning¹ – which in the context of language models means the model develops a broad set of skills and pattern recognition abilities at training time, and then uses those abilities at inference time to rapidly adapt to or recognize the desired task (illustrated in Figure 1.1). Recent work [RWC⁺19] attempts to do this via what we call “in-context learning”, using the text input of a pretrained language model as a form of task specification: the model is conditioned on a natural language instruction and/or a few demonstrations of the task and is then expected to complete further instances of the task simply by predicting what comes next.

While it has shown some initial promise, this approach still achieves results far inferior to fine-tuning – for example [RWC⁺19] achieves only 4% on Natural Questions, and even its 55 F1 CoQa result is now more than 35 points behind the state of the art. Meta-learning clearly requires substantial improvement in order to be viable as a practical method of solving language tasks.

Another recent trend in language modeling may offer a way forward. In recent years the capacity of transformer language models has increased substantially, from 100 million parameters [RNSS18], to 300 million parameters [DCLT18], to 1.5 billion parameters [RWC⁺19], to 8 billion parameters [SPP⁺19], 11 billion parameters [RSR⁺19], and finally 17 billion parameters [Tur20]. Each increase has brought improvements in text synthesis and/or downstream NLP tasks, and there is evidence suggesting that log loss, which correlates well with many downstream tasks, follows a smooth trend of improvement with scale [KMH⁺20]. Since in-context learning involves absorbing many skills and tasks within the parameters of the model, it is plausible that in-context learning abilities might show similarly strong gains with scale.

¹In the context of language models this has sometimes been called “zero-shot transfer”, but this term is potentially ambiguous: the method is “zero-shot” in the sense that no gradient updates are performed, but it often involves providing inference-time demonstrations to the model, so is not truly learning from zero examples. To avoid this confusion, we use the term “meta-learning” to capture the inner-loop / outer-loop structure of the general method, and the term “in context-learning” to refer to the inner loop of meta-learning. We further specialize the description to “zero-shot”, “one-shot”, or “few-shot” depending on how many demonstrations are provided at inference time. These terms are intended to remain agnostic on the question of whether the model learns new tasks from scratch at inference time or simply recognizes patterns seen during training – this is an important issue which we discuss later in the paper, but “meta-learning” is intended to encompass both possibilities, and simply describes the inner-outer loop structure.

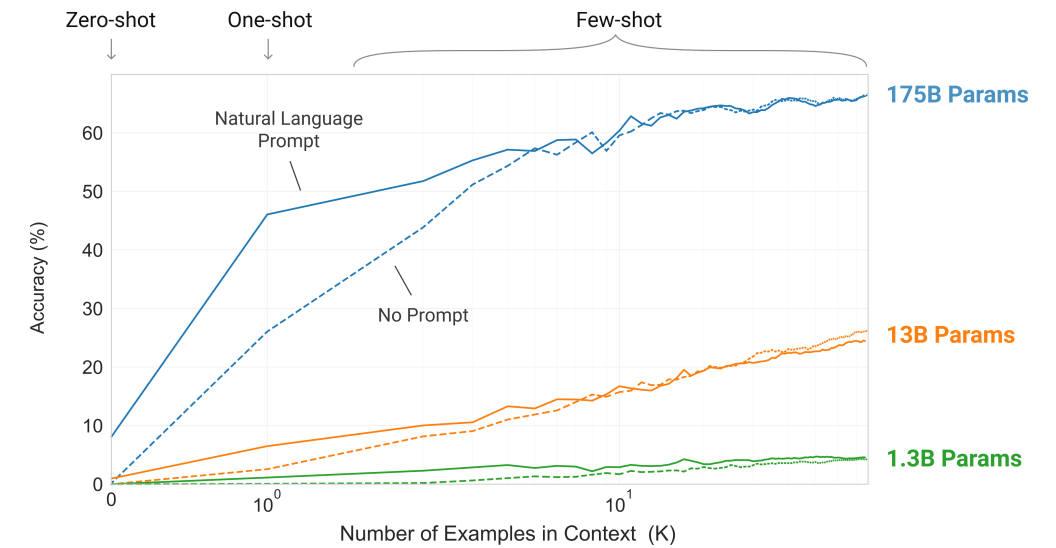


图1.2: 更大的模型对上下文信息的利用效率越来越高。 我们在一个简单的任务上展示了上下文学习性能，该任务要求模型从单词中移除随机符号，包括使用自然语言任务描述和不使用自然语言任务描述的情况（参见第 3.9.2 节）。3.9.2）。大型模型的“上下文学习曲线”更陡峭，表明它们从上下文信息中学习任务的能力有所提高。我们在各种任务中观察到类似的定性行为。

人类能够至少达到合理程度的胜任新任务。除了指出我们当前自然语言处理技术中的一个概念性局限外，这种适应性还有实际优势——它允许人类能够无缝地混合或切换多种任务和技能，例如在长时间的对话中进行加法运算。为了广泛地发挥作用，我们希望有一天我们的自然语言处理系统能够拥有同样的灵活性和通用性。

解决这些问题的一个潜在途径是元学习¹——在语言模型的上下文中，这意味着模型在训练时发展出一套广泛的技术和模式识别能力，然后在推理时使用这些能力快速适应或识别所需任务（如图1.1所示）。1.1）。最近的工作 [RWC⁺19] 试图通过我们称为“上下文学习”的方法来实现这一点，即使用预训练语言模型的文本输入作为任务规范的一种形式：模型在自然语言指令和/或任务的一些示例上进行条件化，然后只需预测接下来会发生什么即可完成进一步实例。

虽然它已展现出一些初步的潜力，但这种方法仍然取得的成果远不如微调——例如[RWC⁺19] 在自然问题上的得分仅为4%，即使其55的F1 CoQa结果也落后于当前最佳水平超过35分。元学习显然需要大幅改进才能作为一种实用的语言任务解决方法而可行。

语言建模领域近期又出现一种新趋势，可能为未来发展提供方向。近年来，Transformer语言模型的容量大幅提升，从1亿参数 [RNSS18]，到3亿参数[DCLT18]，到15亿参数 [RWC⁺19]，到80亿参数 [SPP⁺19]，110亿参数 [RSR⁺19]，最终达到170亿参数 [Tur20]。每次提升都带来了文本合成和/或下游NLP任务的改进，且有证据表明，与许多下游任务高度相关的对数损失，随着规模的增长呈现出平滑的改进趋势 [KMH⁺20]。由于情境学习涉及将许多技能和任务吸收到模型的参数中，因此情境学习能力很可能随着规模的增长同样表现出显著的提升。

¹在语言模型的背景下，这种情况有时被称为“零样本迁移”，但这个术语可能存在歧义：该方法在“没有进行梯度更新”的意义上是“零样本”的，但它通常涉及向模型提供推理时的演示，因此并非真正从零个示例中学习。为了避免这种混淆，我们使用“元学习”一词来概括通用方法的内循环/外循环结构，并使用“上下文学习”一词来指代元学习的内循环。我们进一步根据推理时提供的演示数量将描述专门化为“零样本”、“单样本”或“少样本”。这些术语旨在对模型是否在推理时从零开始学习新任务或仅仅识别训练期间看到的模式保持中立——这是一个我们将在论文后面讨论的重要问题，但“元学习”旨在涵盖这两种可能性，并仅仅描述内-外循环结构。

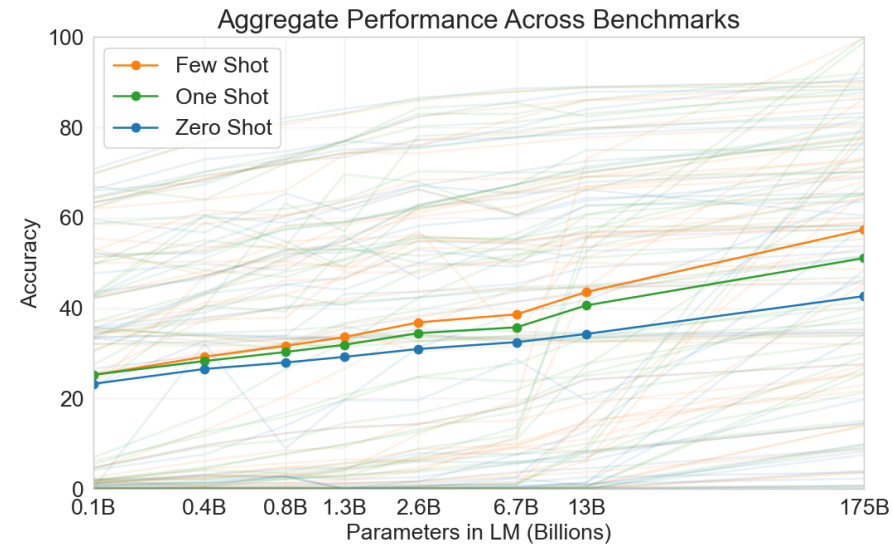


Figure 1.3: Aggregate performance for all 42 accuracy-denominated benchmarks While zero-shot performance improves steadily with model size, few-shot performance increases more rapidly, demonstrating that larger models are more proficient at in-context learning. See Figure 3.8 for a more detailed analysis on SuperGLUE, a standard NLP benchmark suite.

In this paper, we test this hypothesis by training a 175 billion parameter autoregressive language model, which we call GPT-3, and measuring its in-context learning abilities. Specifically, we evaluate GPT-3 on over two dozen NLP datasets, as well as several novel tasks designed to test rapid adaptation to tasks unlikely to be directly contained in the training set. For each task, we evaluate GPT-3 under 3 conditions: (a) “few-shot learning”, or in-context learning where we allow as many demonstrations as will fit into the model’s context window (typically 10 to 100), (b) “one-shot learning”, where we allow only one demonstration, and (c) “zero-shot” learning, where no demonstrations are allowed and only an instruction in natural language is given to the model. GPT-3 could also in principle be evaluated in the traditional fine-tuning setting, but we leave this to future work.

Figure 1.2 illustrates the conditions we study, and shows few-shot learning of a simple task requiring the model to remove extraneous symbols from a word. Model performance improves with the addition of a natural language task description, and with the number of examples in the model’s context, K . Few-shot learning also improves dramatically with model size. Though the results in this case are particularly striking, the general trends with both model size and number of examples in-context hold for most tasks we study. We emphasize that these “learning” curves involve no gradient updates or fine-tuning, just increasing numbers of demonstrations given as conditioning.

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

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

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

A heuristic sense of the overall results can be seen in Figure 1.3, which aggregates the various tasks (though it should not be seen as a rigorous or meaningful benchmark in itself).

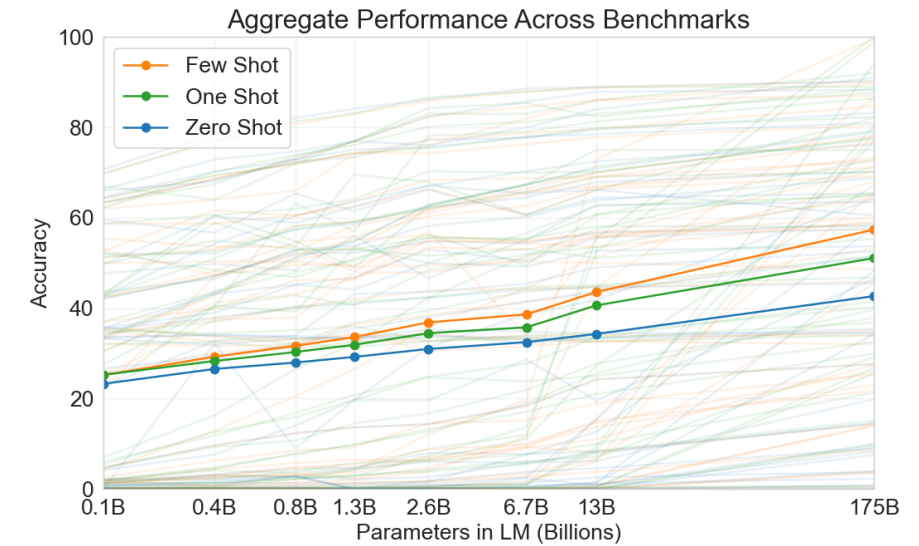


图1.3: 所有42个以准确率衡量的基准测试的聚合性能 虽然零样本学习性能随着模型大小稳步提升, 少样本学习性能增长更快, 这表明更大的模型在情境学习方面更擅长。参见图 3.8, 以获取关于SuperGLUE (一个标准的NLP基准测试套件) 的更详细分析。

在本文中, 我们通过训练一个1750亿参数的自回归语言模型来验证这一假设, 该模型称为GPT-3, 并测量其情境学习能力。具体而言, 我们在超过二十几个NLP数据集上评估GPT-3, 以及设计了几项新任务来测试模型快速适应训练集中不太可能直接包含的任务的能力。对于每个任务, 我们评估GPT-3在三种条件下表现: (a)“少样本学习”, 即情境学习, 我们允许尽可能多的示例填入模型的上上下文窗口 (通常为10到100), (b)“单样本学习”, 我们只允许一个示例, 以及(c)“零样本”学习, 我们不允许任何示例, 仅向模型提供自然语言的指令。原则上, GPT-3也可以在传统微调环境下进行评估, 但我们将其留作未来工作。

图 1.2 展示了我们研究的条件, 并显示了模型需要从单词中移除多余符号的简单任务中的少样本学习。随着自然语言任务描述的添加以及模型上下文中的示例数量, 模型性能得到提升, K 。少样本学习也随着模型大小显著改善。尽管本例中的结果尤为引人注目, 但我们对大多数研究任务而言, 模型大小和上下文中的示例数量都存在的一般趋势都成立。我们强调, 这些“学习”曲线不涉及梯度更新或微调, 只是作为条件提供的示例数量不断增加。

总体而言, 在NLP任务中, GPT-3在零样本和单样本设置中取得了令人鼓舞的结果, 在少样本设置中有时甚至能与或偶尔超越最先进水平 (尽管最先进水平由微调模型保持)。例如, GPT-3在零样本设置中CoQA的F1得分为81.5, 在单样本设置中CoQA的F1得分为84.0, 在少样本设置中F1得分为85.0。同样, GPT-3在零样本设置中TriviaQA的准确率为64.3%, 在单样本设置中为68.0%, 在少样本设置中为71.2%, 后者相对于在同一闭卷设置中运行的微调模型而言是最先进的。

GPT-3在设计用于测试快速适应或即兴推理的任务中也展现出单样本和少样本能力, 这些任务包括解密单词、执行算术运算, 以及在仅见过一次定义后使用新单词造句。我们还表明, 在少样本设置中, GPT-3能够生成合成新闻文章, 人类评估者难以将其与人类生成的文章区分开来。

与此同时, 我们也发现一些任务在GPT-3的规模下少样本性能仍然存在困难。这包括自然语言推理任务 (如ANLI数据集) 和一些阅读理解数据集 (如RACE或QuAC)。通过全面描述GPT-3的优势和劣势, 包括这些局限性, 我们希望激发对语言模型中少样本学习的研究, 并引起对最需要进步领域的关注。

可以从图1.3中看到整体结果的启发式感知, 该图汇总了各种任务 (尽管它不应被视为一个严格或具有意义的基准本身)。

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

In addition to all the above, we also train a series of smaller models (ranging from 125 million parameters to 13 billion parameters) in order to compare their performance to GPT-3 in the zero, one and few-shot settings. Broadly, for most tasks we find relatively smooth scaling with model capacity in all three settings; one notable pattern is that the gap between zero-, one-, and few-shot performance often grows with model capacity, perhaps suggesting that larger models are more proficient meta-learners.

Finally, given the broad spectrum of capabilities displayed by GPT-3, we discuss concerns about bias, fairness, and broader societal impacts, and attempt a preliminary analysis of GPT-3’s characteristics in this regard.

The remainder of this paper is organized as follows. In Section 2, we describe our approach and methods for training GPT-3 and evaluating it. Section 3 presents results on the full range of tasks in the zero-, one- and few-shot settings. Section 4 addresses questions of data contamination (train-test overlap). Section 5 discusses limitations of GPT-3. Section 6 discusses broader impacts. Section 7 reviews related work and Section 8 concludes.

2 Approach

Our basic pre-training approach, including model, data, and training, is similar to the process described in [RWC⁺19], with relatively straightforward scaling up of the model size, dataset size and diversity, and length of training. Our use of in-context learning is also similar to [RWC⁺19], but in this work we systematically explore different settings for learning within the context. Therefore, we start this section by explicitly defining and contrasting the different settings that we will be evaluating GPT-3 on or could in principle evaluate GPT-3 on. These settings can be seen as lying on a spectrum of how much task-specific data they tend to rely on. Specifically, we can identify at least four points on this spectrum (see Figure 2.1 for an illustration):

- **Fine-Tuning (FT)** has been the most common approach in recent years, and involves updating the weights of a pre-trained model by training on a supervised dataset specific to the desired task. Typically thousands to hundreds of thousands of labeled examples are used. The main advantage of fine-tuning is strong performance on many benchmarks. The main disadvantages are the need for a new large dataset for every task, the potential for poor generalization out-of-distribution [MPL19], and the potential to exploit spurious features of the training data [GSL⁺18, NK19], potentially resulting in an unfair comparison with human performance. In this work we do not fine-tune GPT-3 because our focus is on task-agnostic performance, but GPT-3 can be fine-tuned in principle and this is a promising direction for future work.
- **Few-Shot (FS)** is the term we will use in this work to refer to the setting where the model is given a few demonstrations of the task at inference time as conditioning [RWC⁺19], but no weight updates are allowed. As shown in Figure 2.1, for a typical dataset an example has a context and a desired completion (for example an English sentence and the French translation), and few-shot works by giving K examples of context and completion, and then one final example of context, with the model expected to provide the completion. We typically set K in the range of 10 to 100 as this is how many examples can fit in the model’s context window ($n_{\text{ctx}} = 2048$). The main advantages of few-shot are a major reduction in the need for task-specific data and reduced potential to learn an overly narrow distribution from a large but narrow fine-tuning dataset. The main disadvantage is that results from this method have so far been much worse than state-of-the-art fine-tuned models. Also, a small amount of task specific data is still required. As indicated by the name, few-shot learning as described here for language models is related to few-shot learning as used in other contexts in ML [HYC01, VBL⁺16] – both involve learning based on a broad distribution of tasks (in this case implicit in the pre-training data) and then rapidly adapting to a new task.
- **One-Shot (1S)** is the same as few-shot except that only one demonstration is allowed, in addition to a natural language description of the task, as shown in Figure 1. The reason to distinguish one-shot from few-shot and zero-shot (below) is that it most closely matches the way in which some tasks are communicated to humans. For example, when asking humans to generate a dataset on a human worker service (for example Mechanical Turk), it is common to give one demonstration of the task. By contrast it is sometimes difficult to communicate the content or format of a task if no examples are given.

我们还对“数据污染”进行了系统研究——这是一个在Common Crawl等数据集上训练大容量模型时日益严重的问题，因为这些数据集可能因为此类内容通常存在于网络上而包含测试数据集的内容。在本文中，我们开发了系统工具来测量数据污染并量化其扭曲效应。尽管我们发现数据污染对GPT-3在大多数数据集上的性能影响很小，但我们确实发现有几个数据集可能存在结果被夸大的情况，我们根据严重程度要么不报告这些数据集的结果，要么用星号注明。

此外，我们还训练了一系列较小的模型（参数规模从1.25亿到130亿不等），以便在零样本、单样本和少样本设置下与GPT-3的性能进行比较。总体而言，在三个设置中，我们发现大多数任务随着模型容量的增加表现出相对平滑的扩展；一个值得注意的模式是，零样本、单样本和少样本性能之间的差距往往随着模型容量的增加而扩大，这可能表明更大的模型是更擅长的元学习者。

最后，鉴于GPT-3展现出广泛的能力，我们讨论了关于偏见、公平性和更广泛的社会影响方面的担忧，并尝试对GPT-3在这方面的一些特征进行初步分析。

本文的其余部分安排如下。第2节，我们描述了训练GPT-3和评估它的方法。第3节展示了在零样本、单样本和少样本设置下所有任务的结果。第4节探讨了数据污染（训练-测试重叠）的问题。第5节讨论了GPT-3的局限性。第6节讨论了更广泛的影响。第7节回顾了相关工作，第8节总结。

2 方法

我们的基本预训练方法，包括模型、数据和训练，与 [RWC⁺19]中描述的过程相似，模型大小、数据集大小和多样性以及训练长度的扩展相对直接。我们使用情境学习的做法也与 [RWC⁺19]相似，但在这项工作中，我们系统地探索了情境内学习的不同设置。因此，我们从明确定义和对比我们将要在其上评估 GPT-3 或原则上可以在其上评估 GPT-3 的不同设置开始本节。这些设置可以看作是它们倾向于依赖多少特定任务数据的范围上的点。具体来说，我们可以在该范围内至少识别出四个点（见图 2.1 所示）：

- **微调 (FT)** 已成为近年来最常见的方法，涉及通过在特定于所需任务的监督数据集上训练来更新预训练模型的权重。通常使用数千到数百万个标记示例。微调的主要优点是在许多基准测试上表现出色。主要缺点是每个任务都需要一个新的、大型的数据集，以及泛化分布外性能可能较差 [MPL19]，以及利用训练数据中的虚假特征 [GSL⁺18, NK19]，这可能导致与人类性能的不公平比较。在这项工作中我们不微调 GPT-3，因为我们的重点是任务无关的性能，但原则上 GPT-3 可以被微调，这为未来的工作提供了一个有前景的方向。
- **少样本 (FS)** 是我们在本研究中用来指代在推理时向模型提供少量任务示例作为条件，但允许进行权重更新的设置 [RWC⁺19]。如图2.1所示，对于典型数据集，一个示例包含上下文和期望的补全（例如英语句子和法语翻译），而少样本通过提供上下文和补全的示例，然后给出一个最终的上下文示例，模型被期望提供补全。我们通常将 K 设置为10到100之间，因为这是模型上下文窗口中可以容纳的示例数量（ $K \text{ ctx } n$ ，= 2048）。少样本的主要优点是大幅减少对特定任务数据的需要，并减少从大型但狭窄的微调数据集中学习过于狭窄分布的潜在可能性。主要缺点是到目前为止，这种方法的结果远不如最先进微调模型。此外，仍然需要少量特定任务数据。正如其名称所示，这里为语言模型描述的少样本学习与其他上下文中使用的机器学习中的少样本学习 [HYC01, VBL⁺16] 都涉及基于广泛任务分布（在这种情况下隐含于预训练数据中）的学习，然后快速适应新任务
- **单样本学习 (1S)** 与少样本学习相同，只是允许只有一个演示，除了任务的自然语言描述，如图1所示。之所以区分单样本学习与少样本学习及零样本学习（如下所述）的原因是，它最接近某些任务向人类传达的方式。例如，当要求人类在一个人类工作者服务（例如Mechanical Turk）上生成数据集时，通常会给一个任务的演示。相比之下，如果没有给出示例，有时很难传达任务的内容或格式。

The three settings we explore for in-context learning

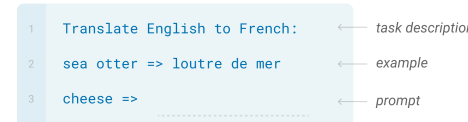
Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.



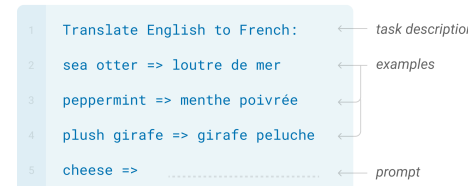
One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.



Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.



Traditional fine-tuning (not used for GPT-3)

Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.



The three settings we explore for in-context learning

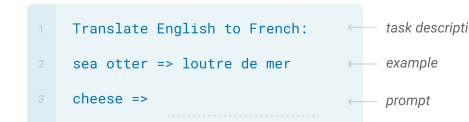
Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.



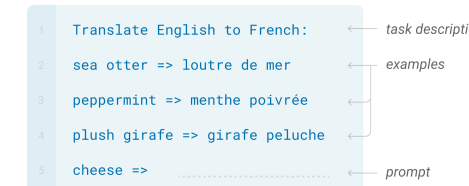
One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.



Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.



Traditional fine-tuning (not used for GPT-3)

Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.



Figure 2.1: Zero-shot, one-shot and few-shot, contrasted with traditional fine-tuning. The panels above show four methods for performing a task with a language model – fine-tuning is the traditional method, whereas zero-, one-, and few-shot, which we study in this work, require the model to perform the task with only forward passes at test time. We typically present the model with a few dozen examples in the few shot setting. Exact phrasings for all task descriptions, examples and prompts can be found in Appendix G.

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

Figure 2.1 shows the four methods using the example of translating English to French. In this paper we focus on zero-shot, one-shot and few-shot, with the aim of comparing them not as competing alternatives, but as different problem settings which offer a varying trade-off between performance on specific benchmarks and sample efficiency. We especially highlight the few-shot results as many of them are only slightly behind state-of-the-art fine-tuned models. Ultimately, however, one-shot, or even sometimes zero-shot, seem like the fairest comparisons to human performance, and are important targets for future work.

Sections 2.1-2.3 below give details on our models, training data, and training process respectively. Section 2.4 discusses the details of how we do few-shot, one-shot, and zero-shot evaluations.

图2.1: 零样本、单样本和少样本, 与传统微调进行对比。 上图展示了使用语言模型执行任务的四种方法——微调是传统方法, 而零样本、单样本和少样本 (我们在此工作中研究的方法) 则要求模型在测试时仅通过前向传递来执行任务。在少样本设置中, 我们通常向模型展示几十个示例。所有任务描述、示例和提示的精确措辞均可在附录G中找到。

- **零样本学习(0S)** 与单样本学习相同, 只是不允许提供演示, 模型仅获得描述任务的自然语言指令。这种方法提供了最大的便利性、鲁棒性潜力以及避免虚假关联 (除非它们在预训练数据的大型语料库中非常普遍), 但也最具挑战性。在某些情况下, 即使没有先前的示例, 人类也可能难以理解任务的格式, 因此这种设置在某些情况下是“不公平地困难”。例如, 如果有人被要求“制作200米短跑世界纪录表”, 这个请求可能是模糊的, 因为它可能不明确表格应该具有的格式或应该包含什么 (即使经过仔细澄清, 精确理解所要求的内容也可能很困难)。尽管如此, 对于至少某些设置, 零样本学习最接近人类执行任务的方式——例如, 在图2.1中的翻译示例中, 人类很可能仅从文本指令中就知道该做什么。

图2.1 展示了使用英语到法语翻译的示例来展示这四种方法。在本文中, 我们重点关注零样本、单样本和少样本, 旨在将它们不是作为竞争性替代方案, 而是作为提供不同权衡 (即在特定基准测试上的性能和样本效率之间) 的问题设置进行比较。我们特别强调少样本结果, 因为其中许多结果仅略逊于最先进微调模型。然而, 最终, 单样本, 甚至有时是零样本, 似乎是与人类性能最公平的比较, 也是未来工作的重要目标。

章节 2.1-2.3 以下将分别介绍我们的模型、训练数据和训练过程。章节 2.4 讨论了我们如何进行少样本、单样本和零样本评估的细节。

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

Table 2.1: Sizes, architectures, and learning hyper-parameters (batch size in tokens and learning rate) of the models which we trained. All models were trained for a total of 300 billion tokens.

2.1 Model and Architectures

We use the same model and architecture as GPT-2 [RWC⁺19], including the modified initialization, pre-normalization, and reversible tokenization described therein, with the exception that we use alternating dense and locally banded sparse attention patterns in the layers of the transformer, similar to the Sparse Transformer [CGRS19]. To study the dependence of ML performance on model size, we train 8 different sizes of model, ranging over three orders of magnitude from 125 million parameters to 175 billion parameters, with the last being the model we call GPT-3. Previous work [KMH⁺20] suggests that with enough training data, scaling of validation loss should be approximately a smooth power law as a function of size; training models of many different sizes allows us to test this hypothesis both for validation loss and for downstream language tasks.

Table 2.1 shows the sizes and architectures of our 8 models. Here n_{params} is the total number of trainable parameters, n_{layers} is the total number of layers, d_{model} is the number of units in each bottleneck layer (we always have the feedforward layer four times the size of the bottleneck layer, $d_{\text{ff}} = 4 * d_{\text{model}}$), and d_{head} is the dimension of each attention head. All models use a context window of $n_{\text{ctx}} = 2048$ tokens. We partition the model across GPUs along both the depth and width dimension in order to minimize data-transfer between nodes. The precise architectural parameters for each model are chosen based on computational efficiency and load-balancing in the layout of models across GPU's. Previous work [KMH⁺20] suggests that validation loss is not strongly sensitive to these parameters within a reasonably broad range.

2.2 Training Dataset

Datasets for language models have rapidly expanded, culminating in the Common Crawl dataset² [RSR⁺19] constituting nearly a trillion words. This size of dataset is sufficient to train our largest models without ever updating on the same sequence twice. However, we have found that unfiltered or lightly filtered versions of Common Crawl tend to have lower quality than more curated datasets. Therefore, we took 3 steps to improve the average quality of our datasets: (1) we downloaded and filtered a version of CommonCrawl based on similarity to a range of high-quality reference corpora, (2) we performed fuzzy deduplication at the document level, within and across datasets, to prevent redundancy and preserve the integrity of our held-out validation set as an accurate measure of overfitting, and (3) we also added known high-quality reference corpora to the training mix to augment CommonCrawl and increase its diversity.

Details of the first two points (processing of Common Crawl) are described in Appendix A. For the third, we added several curated high-quality datasets, including an expanded version of the WebText dataset [RWC⁺19], collected by scraping links over a longer period of time, and first described in [KMH⁺20], two internet-based books corpora (Books1 and Books2) and English-language Wikipedia.

Table 2.2 shows the final mixture of datasets that we used in training. The CommonCrawl data was downloaded from 41 shards of monthly CommonCrawl covering 2016 to 2019, constituting 45TB of compressed plaintext before filtering and 570GB after filtering, roughly equivalent to 400 billion byte-pair-encoded tokens. Note that during training, datasets are not sampled in proportion to their size, but rather datasets we view as higher-quality are sampled more frequently, such that CommonCrawl and Books2 datasets are sampled less than once during training, but the other datasets are sampled 2-3 times. This essentially accepts a small amount of overfitting in exchange for higher quality training data.

²<https://commoncrawl.org/the-data/>

模型名称	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	批大小	学习率
GPT-3 小型	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 中型	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 大型	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 27亿	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 67亿	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 130亿	130亿	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B 或 "GPT-3"	1750亿	96	12288	96	128	3.2M	0.6×10^{-4}

表2.1: 所训练模型的尺寸、架构和学习超参数（批大小以token计和学习率）。所有模型总计训练了3000亿token。

2.1 模型与架构

我们使用与 GPT-2 [RWC⁺19]相同的模型和架构，包括其中描述的修改初始化、预归一化和可逆分词，唯一的区别是我们使用交替的密集和局部带状稀疏注意力模式在 Transformer 的层中，类似于稀疏 Transformer [CGRS19]。为了研究 ML 性能与模型大小的依赖关系，我们训练了 8 个不同大小的模型，参数量跨越三个数量级，从 1.25 亿到 1750 亿，最后一个是我们称为 GPT-3 的模型。先前工作 [KMH⁺20]表明，在有足够训练数据的情况下，验证损失的缩放应该近似于一个平滑的幂律函数，作为大小的函数；训练许多不同大小的模型使我们能够针对验证损失和下游语言任务测试这一假设。

表2.1 显示了我们8个模型的尺寸和架构。这里 n_{params} 是可训练参数的总数 n_{layers} 是层数的总数 d_{model} 是每个瓶颈层的单元数（我们始终让前馈层的尺寸是瓶颈层的四倍 $d_{\text{ff}} = 4 * d_{\text{model}}$ ）， d_{head} 是每个注意力头的维度。所有模型使用 $n_{\text{ctx}} = 2048$ token 的上下文窗口。我们沿深度和宽度维度在GPU之间划分模型，以最小化节点间的数据传输。每个模型的精确架构参数基于计算效率和模型在GPU布局中的负载均衡来选择。先前工作 [KMH⁺20] 表明，在合理的宽泛范围内，验证损失对这些参数并不敏感。

2.2 训练数据集

用于语言模型的训练数据集已迅速扩展，最终形成了CommonCrawl数据集² [RSR⁺19] 其包含近万亿个单词。这种规模的数据集足以训练我们最大的模型，且无需在同一个序列上重复更新。然而，我们发现未经过滤或轻度过滤的Common Crawl版本，其质量通常低于更经过精心策划的数据集。因此，我们采取了三个步骤来提高我们数据集的平均质量：(1) 我们下载并过滤了一个与一系列高质量参考语料库相似度的CommonCrawl版本，(2) 我们在数据集内部和之间执行了文档级别的模糊去重，以防止冗余并保持我们保留的验证集的完整性，将其作为衡量过拟合的准确指标，(3) 我们还添加了已知的高质量参考语料库到训练组合中，以增强CommonCrawl并增加其多样性。

前两点（处理Common Crawl）的细节在附录A中描述。对于第三点，我们添加了几个精心策划的高质量数据集，包括WebText数据集的扩展版本 [RWC⁺19]，该数据集通过更长时间段内抓取链接收集，并在 [KMH⁺20]中首次描述，还包括两个基于互联网的书籍语料库（Books1和Books2）以及英语维基百科。

表 2.2 展示了我们在训练中使用的最终数据集混合。CommonCrawl 数据从 2016 年至 2019 年的 41 个月度 CommonCrawl 分片中下载，构成 45TB 的压缩明文数据，过滤后为 570GB，约相当于 4000 亿字节对编码 token。请注意，在训练过程中，数据集并非按其大小比例采样，而是我们视为高质量的数据集采样频率更高，因此 CommonCrawl 和 Books2 数据集在训练中采样不足一次，而其他数据集采样 2-3 次。这本质上是以少量过拟合为代价，换取了更高质量的训练数据。

²<https://commoncrawl.org/the-data/>

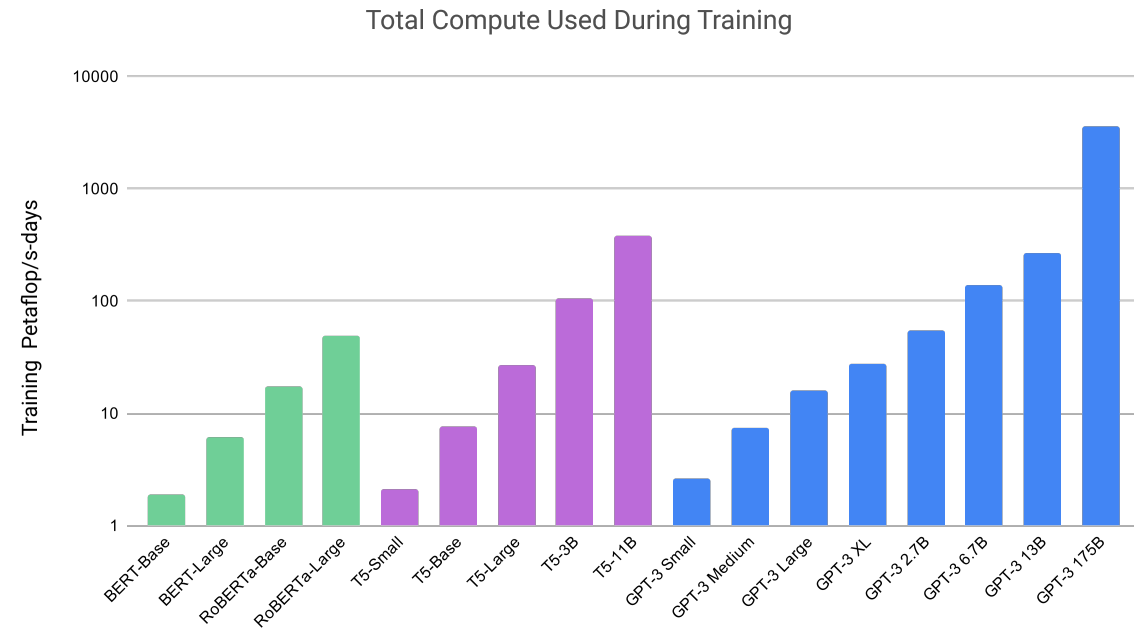


Figure 2.2: Total compute used during training. Based on the analysis in Scaling Laws For Neural Language Models [KMH⁺20] we train much larger models on many fewer tokens than is typical. As a consequence, although GPT-3 3B is almost 10x larger than RoBERTa-Large (355M params), both models took roughly 50 petaflop/s-days of compute during pre-training. Methodology for these calculations can be found in Appendix D.

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

Table 2.2: Datasets used to train GPT-3. “Weight in training mix” refers to the fraction of examples during training that are drawn from a given dataset, which we intentionally do not make proportional to the size of the dataset. As a result, when we train for 300 billion tokens, some datasets are seen up to 3.4 times during training while other datasets are seen less than once.

A major methodological concern with language models pretrained on a broad swath of internet data, particularly large models with the capacity to memorize vast amounts of content, is potential contamination of downstream tasks by having their test or development sets inadvertently seen during pre-training. To reduce such contamination, we searched for and attempted to remove any overlaps with the development and test sets of all benchmarks studied in this paper. Unfortunately, a bug in the filtering caused us to ignore some overlaps, and due to the cost of training it was not feasible to retrain the model. In Section 4 we characterize the impact of the remaining overlaps, and in future work we will more aggressively remove data contamination.

2.3 Training Process

As found in [KMH⁺20, MKAT18], larger models can typically use a larger batch size, but require a smaller learning rate. We measure the gradient noise scale during training and use it to guide our choice of batch size [MKAT18]. Table 2.1 shows the parameter settings we used. To train the larger models without running out of memory, we use a mixture of model parallelism within each matrix multiply and model parallelism across the layers of the network. All models were trained on V100 GPU’s on part of a high-bandwidth cluster provided by Microsoft. Details of the training process and hyperparameter settings are described in Appendix B.

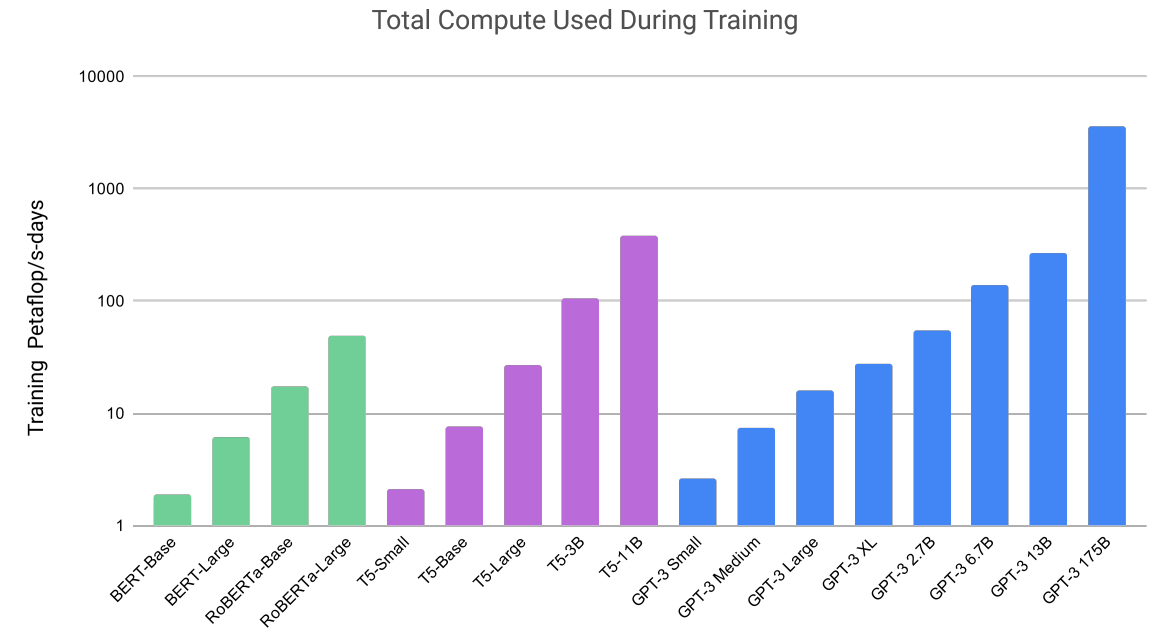


图 2.2: 训练期间使用的总计算资源。根据《神经语言模型的扩展定律》中的分析[KMH⁺20]我们在许多更少的 token 上训练了远比典型情况大得多的模型。因此，尽管 GPT-3 3B 比 RoBERTa-Large (355M 参数) 大近 10 倍，但这两个模型在预训练期间都大致使用了 50 petaflop/s-天的计算资源。这些计算的方法可以在附录D. 中找到。

数据集	数量 (token)	权重在混合	经过的轮数当训练300B个token时
Common Crawl (过滤后)	4100亿	60%	0.44
WebText2	190亿	22%	2.9
书籍1	120亿	8%	1.9
书籍2	550亿	8%	0.43
维基百科	30亿	3%	3.4

表2.2: 用于训练GPT-3的数据集。“训练混合中的权重”指的是在训练过程中从给定数据集中抽取的示例的比例，我们有意不使其与数据集的大小成比例。因此，当我们训练3000亿个token时，某些数据集在训练过程中会被看到高达3.4次，而其他数据集则少于一次。

使用广泛互联网数据预训练的语言模型，尤其是能够记忆海量内容的大型模型，存在一个主要的方法论问题：其测试或开发集可能在预训练期间无意中被看到，从而污染下游任务。为减少此类污染，我们搜索并尝试移除与本文所研究所有基准测试的开发和测试集的重叠。不幸的是，过滤过程中的一个错误导致我们忽略了一些重叠，且由于训练成本高，重新训练模型不可行。在 4 部分，我们描述了剩余重叠的影响，并在未来工作中将更积极地移除数据污染。

2.3 训练过程

如 [KMH⁺20 中所发现, MKAT18], 大型模型通常可以使用更大的批大小，但需要更小的学习率。我们在训练过程中测量梯度噪声尺度，并以此指导我们选择批大小 [MKAT18]。表 2.1 展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大型模型，我们使用矩阵乘法内的模型并行和跨网络层的模型并行。所有模型都在微软提供的高带宽集群的 V100 GPU 上进行训练。训练过程和超参数设置的详细信息在附录 B 中描述。

2.4 Evaluation

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. For Winograd (the original, not SuperGLUE version) there is only one dataset, so we draw conditioning examples directly from it.

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

On tasks that involve choosing one correct completion from several options (multiple choice), we provide K examples of context plus correct completion, followed by one example of context only, and compare the LM likelihood of each completion. For most tasks we compare the per-token likelihood (to normalize for length), however on a small number of datasets (ARC, OpenBookQA, and RACE) we gain additional benefit as measured on the development set by normalizing by the unconditional probability of each completion, by computing $\frac{P(\text{completion}|\text{context})}{P(\text{completion}|\text{answer_context})}$, where `answer_context` is the string "Answer: " or "A: " and is used to prompt that the completion should be an answer but is otherwise generic.

On tasks that involve binary classification, we give the options more semantically meaningful names (e.g. “True” or “False” rather than 0 or 1) and then treat the task like multiple choice; we also sometimes frame the task similar to what is done by [RSR+19] (see Appendix G) for details.

On tasks with free-form completion, we use beam search with the same parameters as [RSR+19]: a beam width of 4 and a length penalty of $\alpha = 0.6$. We score the model using F1 similarity score, BLEU, or exact match, depending on what is standard for the dataset at hand.

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. We do submit to the test server on a small number of datasets (SuperGLUE, TriviaQA, PiQa) where we were able to make submission work, and we submit only the 200B few-shot results, and report development set results for everything else.

3 Results

In Figure 3.1 we display training curves for the 8 models described in Section 2. For this graph we also include 6 additional extra-small models with as few as 100,000 parameters. As observed in [KMH+20], language modeling performance follows a power-law when making efficient use of training compute. After extending this trend by two more orders of magnitude, we observe only a slight (if any) departure from the power-law. One might worry that these improvements in cross-entropy loss come only from modeling spurious details of our training corpus. However, we will see in the following sections that improvements in cross-entropy loss lead to consistent performance gains across a broad spectrum of natural language tasks.

Below, we evaluate the 8 models described in Section 2 (the 175 billion parameter parameter GPT-3 and 7 smaller models) on a wide range of datasets. We group the datasets into 9 categories representing roughly similar tasks.

In Section 3.1 we evaluate on traditional language modeling tasks and tasks that are similar to language modeling, such as Cloze tasks and sentence/paragraph completion tasks. In Section 3.2 we evaluate on “closed book” question answering tasks: tasks which require using the information stored in the model’s parameters to answer general knowledge questions. In Section 3.3 we evaluate the model’s ability to translate between languages (especially one-shot and few-shot). In Section 3.4 we evaluate the model’s performance on Winograd Schema-like tasks. In Section 3.5 we evaluate on datasets that involve commonsense reasoning or question answering. In Section 3.6 we evaluate on reading comprehension tasks, in Section 3.7 we evaluate on the SuperGLUE benchmark suite, and in 3.8 we briefly explore NLI. Finally, in Section 3.9, we invent some additional tasks designed especially to probe in-context learning abilities – these tasks focus on on-the-fly reasoning, adaptation skills, or open-ended text synthesis. We evaluate all tasks in the few-shot, one-shot, and zero-shot settings.

2.4 评估

对于少样本学习，我们通过从该任务的训练集中随机抽取 K 个示例作为条件，并根据任务的不同使用 1 或 2 换行符进行分隔，来评估评估集中的每个示例。对于 LAMBADA 和 StoryCloze 没有可用的监督训练集，因此我们从开发集抽取条件示例，并在测试集上评估。对于 Winograd（原始版本，非 SuperGLUE 版本）只有一个数据集，因此我们直接从该数据集中抽取条件示例。

K 可以是 0 到模型上下文窗口允许的最大值之间的任何值，该上下文窗口 $n_{\text{ctx}} = 2048$ 对所有模型都适用，通常适合 10 到 100 个示例。较大的 K 值通常更好，但不一定总是更好，因此当有单独的开发集和测试集时，我们在开发集上尝试几个 K 的值，然后在测试集上运行最佳值。对于某些任务（见附录 G），我们还使用自然语言提示，作为（或用于 $K = 0$ ，而不是）演示。

对于需要从多个选项选择一个正确补全的任务（多项选择），我们提供 K 上下文示例加上正确补全，然后是一个只有上下文的示例，并比较每个补全的语言模型可能性。对于大多数任务，我们比较每个 token 的可能性（以归一化长度），但在少数数据集（ARC、OpenBookQA 和 RACE）上，通过归一化每个补全的无条件概率，在开发集上获得额外的好处，通过计算 $\frac{P(\text{补全}|\text{上下文})}{P(\text{补全}|\text{答案上下文})}$ ，其中答案上下文是字符串 "Answer: " 或 "A: "，用于提示补全应该是答案，但除此之外是通用的。

在涉及二元分类的任务中，我们给选项赋予更具语义意义的名称（例如“真”或“假”而不是 0 或 1），然后像多项选择一样处理该任务；我们有时将任务框架化，类似于 [RSR+19]（参见附录 G）获取详细信息。

在涉及自由形式补全的任务中，我们使用与 [RSR+19] 相同的参数的集束搜索：集束宽度为 4，长度惩罚为 $\alpha = 0.6$ 。我们使用 F1 相似度分数、BLEU 或完全匹配对模型进行评分，具体取决于数据集的标准。

当公开可用时，最终结果在每个模型大小和学习设置（零、一和少样本）上报告在测试集上。当测试集是私有时，我们的模型通常太大而无法在测试服务器上运行，因此我们在开发集上报告结果。我们在少量数据集（SuperGLUE、TriviaQA、PiQa）上提交测试服务器，因为我们能够使提交工作，我们只提交 200B 少样本结果，并为其他所有内容报告开发集结果。

3 结果

在图 3.1 中，我们展示了第 2 节中描述的 8 个模型的训练曲线。对于这个图表，我们还包含了 6 个额外的超小型模型，其参数数量少至 100,000 个。如 [KMH+20] 所示，当高效利用训练计算资源时，语言建模性能遵循幂律。在将这一趋势扩展两个数量级后，我们观察到只有微小的（甚至没有）偏离幂律的现象。人们可能会担心，这些交叉熵损失的改进仅仅来自我们对训练语料库中虚假细节的建模。然而，在接下来的几节中，我们将看到交叉熵损失的改进带来了跨广泛自然语言任务的持续性能提升。

下面，我们评估了第 2（包括 1750 亿参数的 GPT-3 和 7 个更小的模型）在一系列数据集上的表现。我们将数据集分为 9 个类别，代表大致相似的任务。

在 3.1 中，我们评估了传统语言建模任务和与语言建模相似的 3.2 任务，例如完形填空和句子/段落补全任务。在 3.3 中，我们评估了模型在跨语言翻译方面的能力（尤其是单样本和少样本）。在 3.4 中，我们评估了模型在 Winograd Schema 类任务上的表现。在 3.5 中，我们评估了涉及常识推理或问答的数据集。在 3.6 中，我们评估了阅读理解任务，在 3.7 中，我们评估了 SuperGLUE 基准测试套件，并在 3.8 中，我们简要探索了 NLI。最后，在 3.9 中，我们设计了一些额外的任务，专门用于测试情境学习能力——这些任务侧重于即时推理、适应能力或开放式文本合成。我们评估所有任务在少样本、单样本和零样本设置下。

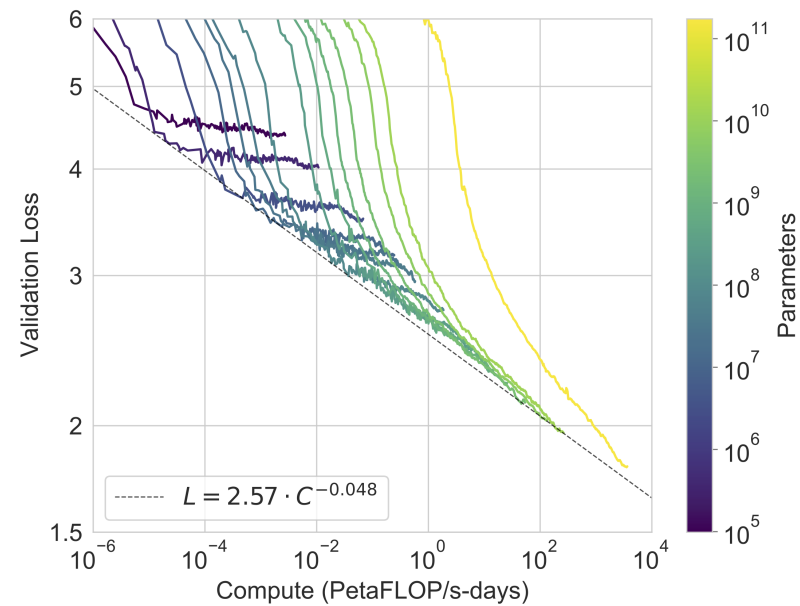


Figure 3.1: Smooth scaling of performance with compute. Performance (measured in terms of cross-entropy validation loss) follows a power-law trend with the amount of compute used for training. The power-law behavior observed in [KMH⁺20] continues for an additional two orders of magnitude with only small deviations from the predicted curve. For this figure, we exclude embedding parameters from compute and parameter counts.

Setting	PTB
SOTA (Zero-Shot)	35.8 ^a
GPT-3 Zero-Shot	20.5

Table 3.1: Zero-shot results on PTB language modeling dataset. Many other common language modeling datasets are omitted because they are derived from Wikipedia or other sources which are included in GPT-3’s training data. ^a[RWC⁺19]

3.1 Language Modeling, Cloze, and Completion Tasks

In this section we test GPT-3’s performance on the traditional task of language modeling, as well as related tasks that involve predicting a single word of interest, completing a sentence or paragraph, or choosing between possible completions of a piece of text.

3.1.1 Language Modeling

We calculate zero-shot perplexity on the Penn Tree Bank (PTB) [MKM⁺94] dataset measured in [RWC⁺19]. We omit the 4 Wikipedia-related tasks in that work because they are entirely contained in our training data, and we also omit the one-billion word benchmark due to a high fraction of the dataset being contained in our training set. PTB escapes these issues due to predating the modern internet. Our largest model sets a new SOTA on PTB by a substantial margin of 15 points, achieving a perplexity of 20.50. Note that since PTB is a traditional language modeling dataset it does not have a clear separation of examples to define one-shot or few-shot evaluation around, so we measure only zero-shot.

3.1.2 LAMBADA

The LAMBADA dataset [PKL⁺16] tests the modeling of long-range dependencies in text – the model is asked to predict the last word of sentences which require reading a paragraph of context. It has recently been suggested that the continued scaling of language models is yielding diminishing returns on this difficult benchmark. [BHT⁺20] reflect on the small 1.5% improvement achieved by a doubling of model size between two recent state of the art results ([SPP⁺19]

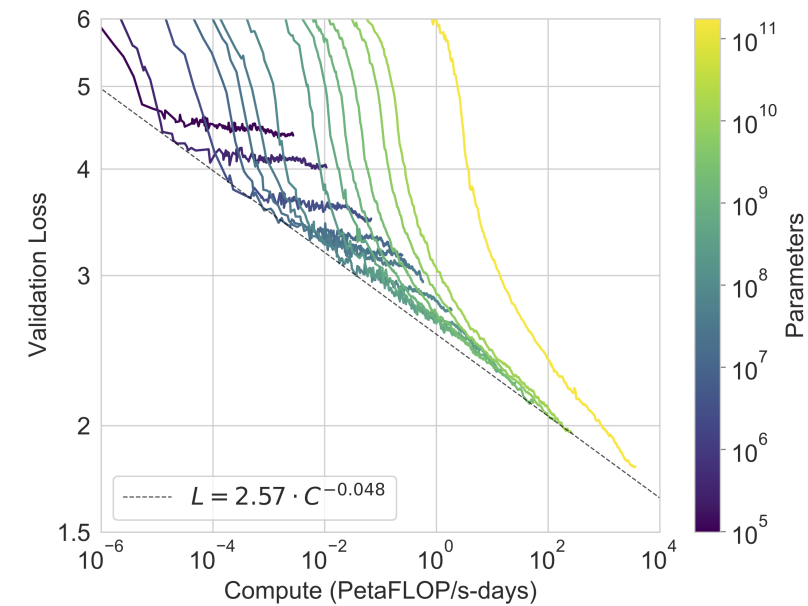


图3.1: 性能随计算资源的平滑扩展。性能（以交叉熵验证损失衡量）随着用于训练的计算资源量呈幂律趋势变化。在KMH⁺20]中观察到的幂律行为，在仅略偏离预测曲线的情况下，继续延伸了两个数量级。对于此图，我们不包括嵌入参数在计算资源和参数计数中。

设置	PTB
SOTA (零样本)	35.8 ^a
GPT-3零样本	20.5

表3.1: 在PTB语言建模数据集上的零样本结果。许多其他常见的语言建模数据集被省略，因为它们源自维基百科或其他包含在GPT-3训练数据中的来源。 ^a[RWC⁺19]

3.1 语言建模、完形填空和补全任务

在本节中，我们测试了GPT-3在传统语言建模任务上的表现，以及涉及预测单个目标词、补全句子或段落，或从文本的可能补全中选择的任务。

3.1.1 语言建模

我们在宾夕法尼亚树银行（PTB）上计算零样本困惑度 [MKM⁺94] 数据集，该数据集以 [RWC⁺19]为度量单位。我们省略了该工作中包含的4个维基百科相关任务，因为它们完全包含在我们的训练数据中，我们还省略了十亿词基准，因为数据集中有很大一部分包含在我们的训练集中。PTB由于早于现代互联网而避免了这些问题。我们的最大模型在PTB上以15分的显著优势创造了新的SOTA，达到了20.50的困惑度。请注意，由于PTB是一个传统的语言建模数据集，它没有明确的示例划分来定义单样本或少样本评估，因此我们仅测量零样本。

3.1.2 LAMBADA

LAMBADA 数据集 [PKL⁺16] 测试文本中远距离依赖的建模——模型被要求预测需要阅读段落上下文的句子的最后一个词。最近有建议称，语言模型的持续扩展在这个困难基准上正在产生递减的回报。 [BHT⁺20] 反思最近两个最先进结果之间模型大小翻倍所取得的1.5%的微小改进 ([SPP⁺19]

Setting	LAMBADA (acc)	LAMBADA (ppl)	StoryCloze (acc)	HellaSwag (acc)
SOTA	68.0 ^a	8.63 ^b	91.8^c	85.6^d
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

Table 3.2: 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]

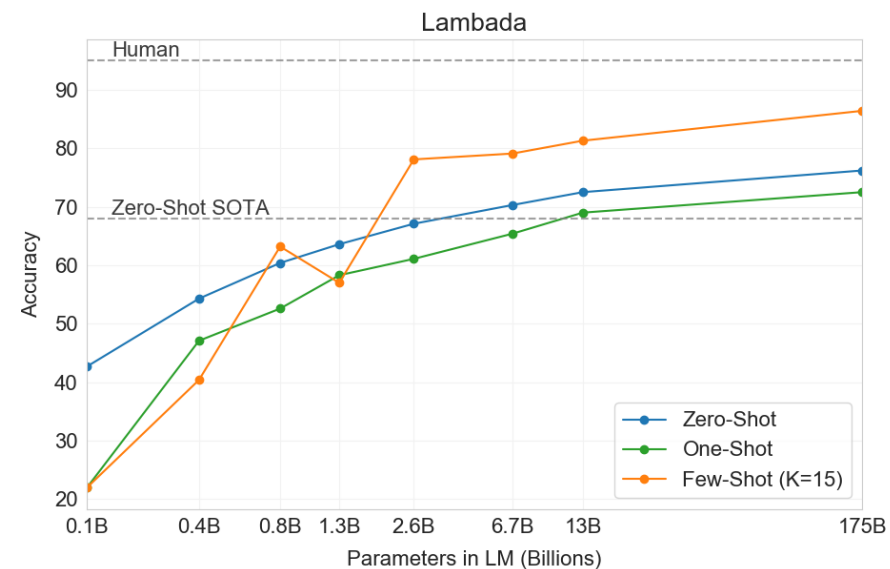


Figure 3.2: On LAMBADA, the few-shot capability of language models results in a strong boost to accuracy. GPT-3 2.7B outperforms the SOTA 17B parameter Turing-NLG [Tur20] in this setting, and GPT-3 175B advances the state of the art by 18%. Note zero-shot uses a different format from one-shot and few-shot as described in the text.

and [Tur20]) and argue that “continuing to expand hardware and data sizes by orders of magnitude is not the path forward”. We find that path is still promising and in a zero-shot setting GPT-3 achieves 76% on LAMBADA, a gain of 8% over the previous state of the art.

LAMBADA is also a demonstration of the flexibility of few-shot learning as it provides a way to address a problem that classically occurs with this dataset. Although the completion in LAMBADA is always the last word in a sentence, a standard language model has no way of knowing this detail. It thus assigns probability not only to the correct ending but also to other valid continuations of the paragraph. This problem has been partially addressed in the past with stop-word filters [RWC⁺19] (which ban “continuation” words). The few-shot setting instead allows us to “frame” the task as a cloze-test and allows the language model to infer from examples that a completion of exactly one word is desired. We use the following fill-in-the-blank format:

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

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

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

设置	LAMBADA (准确率)	LAMBADA (ppl)	StoryCloze (准确率)	HellaSwag (准确率)
SOTA	68.0 ^a	8.63 ^b	91.8^c	85.6^d
GPT-3零样本学习	76.2	3.00	83.2	78.9
GPT-3单样本学习	72.5	3.35	84.7	78.1
GPT-3少样本学习	86.4	1.92	87.7	79.3

表3.2: 在完形填空和补全任务上的性能。 GPT-3显著提升了在LAMBADA上的SOTA表现, 同时在两个困难的补全预测数据集上取得了不错的性能。 ^a [Tur20] ^b[RWC⁺19] ^c[LDL19] ^d[LCH⁺20]

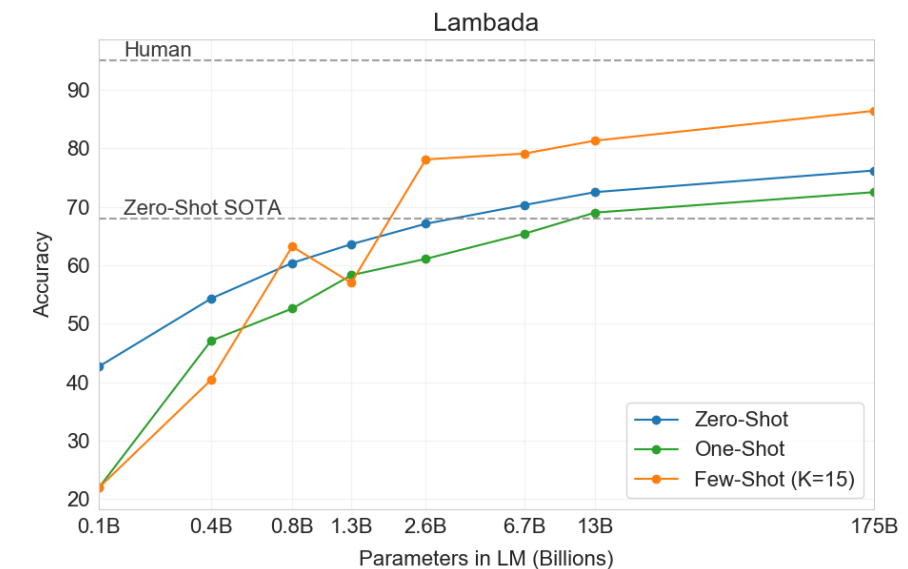


图3.2: 在LAMBADA上, 语言模型的少样本学习能力显著提升了准确率。GPT-3 27亿参数优于SOTA 17亿参数的 Turing-NLG [Tur20] 在当前设置下, GPT-3 175B将SOTA提升了18%。注意零样本学习与单样本学习和少样本学习在文本中描述的格式不同。

和 [Tur20]) 并认为“继续按数量级扩大硬件和数据规模并非前进的道路”。我们发现这条道路仍然充满希望, 在零样本设置下 GPT-3 在 LAMBADA 上达到 76%, 再次领先于之前的最先进结果 8%。

LAMBADA 也是少样本学习灵活性的一个演示, 因为它提供了一种解决这个数据集经典问题的方法。尽管 LAMBADA 中的补全总是出现在句子的最后一个词, 但标准语言模型无法知道这个细节。因此, 它会不仅对正确结尾, 也对段落的其他有效延续分配概率。这个问题过去曾通过停用词过滤器部分地得到解决 [RWC⁺19] (禁止“延续”词的过滤器)。少样本设置则允许我们将任务“框架化”为完形填空测试, 并允许语言模型从示例中推断出期望恰好一个词的补全。我们使用以下填空格式:

爱丽丝和鲍勃是朋友。爱丽丝去看望她的朋友。 → 鲍勃乔治买了一
些棒球装备、一个球、一个手套和 →。

当以这种方式呈现示例时, GPT-3在少样本设置中达到86.4%的准确率, 比之前的顶尖水平提高了18%以上。我们观察到, 少量样本性能随模型大小显著提升。虽然这种设置使最小模型的性能降低了近20%, 但对 GPT-3而言, 它将准确率提高了10%。最后, 填空方法在单样本学习中无效, 其表现始终不如零样本设置。也许这是因为所有模型仍然需要多个示例才能识别模式。

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

Table 3.3: 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.

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

3.1.3 HellaSwag

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

3.1.4 StoryCloze

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

3.2 Closed Book Question Answering

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

The results for GPT-3 are shown in Table 3.3. On TriviaQA, we achieve 64.3% in the zero-shot setting, 68.0% in the one-shot setting, and 71.2% in the few-shot setting. The zero-shot result already outperforms the fine-tuned T5-11B by 14.2%, and also outperforms a version with Q&A tailored span prediction during pre-training by 3.8%. The one-shot result improves by 3.7% and matches the SOTA for an open-domain QA system which not only fine-tunes but also makes use of a learned retrieval mechanism over a 15.3B parameter dense vector index of 21M documents [LPP+20]. GPT-3’s few-shot result further improves performance another 3.2% beyond this.

On WebQuestions (WebQs), GPT-3 achieves 14.4% in the zero-shot setting, 25.3% in the one-shot setting, and 41.5% in the few-shot setting. This compares to 37.4% for fine-tuned T5-11B, and 44.7% for fine-tuned T5-11B+SSM, which uses a Q&A-specific pre-training procedure. GPT-3 in the few-shot setting approaches the performance of state-of-the-art fine-tuned models. Notably, compared to TriviaQA, WebQS shows a much larger gain from zero-shot to few-shot (and indeed its zero-shot and one-shot performance are poor), perhaps suggesting that the WebQs questions

设置	自然问答	WebQS	TriviaQA
RAG (微调, 开放域) [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-3零样本学习	14.6	14.4	64.3
GPT-3单样本学习	23.0	25.3	68.0
GPT-3少样本	29.9	41.5	71.2

表3.3: 在三个开放域QA任务上的结果。 GPT-3在少样本、单样本和零样本设置中显示，与闭卷和开放域设置的前SOTA结果相比。TriviaQA单样本结果在wiki分割测试服务器上评估。

需要注意的是，对测试集污染的分析表明，LAMBADA数据集中有显著一部分似乎存在于我们的训练数据中——然而，第4节的分析表明其对性能的影响可以忽略不计。

3.1.3 HellaSwag

HellaSwag数据集 [ZHB+19] 涉及挑选故事或指令的最佳结局。这些示例经过对抗性挖掘，对语言模型来说很困难，但对人类来说却容易（人类达到了95.6%的准确率）。GPT-3在单样本设置中达到78.1%的准确率，在少样本设置中达到79.3%的准确率，优于微调的1.5B参数语言模型的75.4%的准确率 [ZHR+19]，但仍然比微调的多任务模型ALUM实现的85.6%的整体SOTA低不少。

3.1.4 StoryCloze

我们接下来在StoryCloze 2016数据集上评估GPT-3 [MCH+16]，该数据集涉及为五句长的故事选择正确的结局句。在这里，GPT-3在零样本设置中达到83.2%，在少样本设置中达到87.7%（有 $K = 70$ ）。这仍然比使用BERT模型微调的SOTA低4.1%，但比之前的零样本结果提高了约10%。 [LDL19]。

3.2 闭卷问答

在本节中，我们测量 GPT-3 回答关于广泛事实性知识问题的能力。由于可能的查询量巨大，这项任务通常通过使用信息检索系统来查找相关文本，并结合一个能够根据问题和检索到的文本生成答案的模型来处理。由于这种设置允许系统搜索并基于可能包含答案的文本进行条件化，因此它被标记为“开卷”。 [RRS20]最近，他们证明了一个大语言模型可以直接回答问题，而无需依赖辅助信息。他们将这种更严格的评估设置为“闭卷”。他们的工作表明，即使更高容量的模型也能表现更好，我们用 GPT-3 来验证这一假设。我们在 [RRS20]中的 3 个数据集上评估 GPT-3: 自然问题 [KPR+19], WebQuestions [BCFL13], 和 TriviaQA [JCWZ17], 并使用相同的分割方式。请注意，除了所有结果都在闭卷设置下外，我们使用的少样本、单样本和零样本评估比之前的闭卷问答工作更为严格：除了不允许外部内容外，甚至也不允许在问答数据集本身上进行微调。

GPT-3 的结果如表 3.3所示。在 TriviaQA 上，我们在零样本设置中达到 64.3%，在单样本设置中达到 68.0%，在少样本设置中达到 71.2%。零样本结果已经比微调后的 T5-11B 高 14.2%，同时也比在预训练期间使用问答定制跨度预测的版本高 3.8%。单样本结果提高了 3.7%，并达到了一个开放域问答系统的 SOTA，该系统不仅进行微调，还利用了一个 15.3B 参数的密集向量索引（包含 21M 篇文档）上学习到的检索机制 [LPP+20]。GPT-3 的少样本结果在此基础上进一步提高了 3.2%。

在 WebQuestions (WebQs) 上，GPT-3 在零样本设置中达到 14.4%，在单样本设置中达到 25.3%，在少样本设置中达到 41.5%。这与其他最先进微调模型相比，微调的 T5-11B 达到 37.4%，而使用问答特定预训练流程的微调 T5-11B+SSM 达到 44.7%。在少样本设置中，GPT-3 的性能接近最先进微调模型的性能。值得注意的是，与 TriviaQA 相比，WebQS 从零样本到少样本的增益要大得多（而且它的零样本和单样本性能确实很差），这可能表明 WebQs 的问题

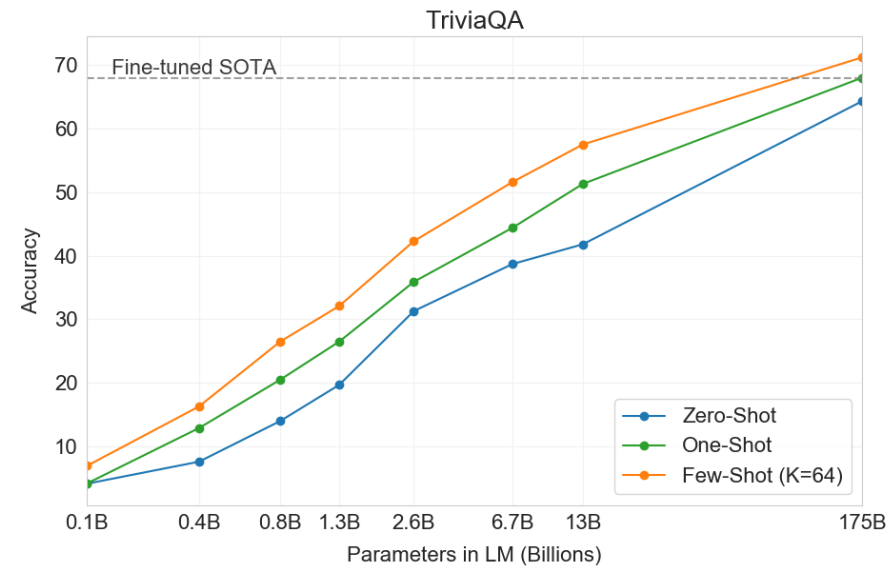


Figure 3.3: On TriviaQA GPT-3’s performance grows smoothly with model size, suggesting that language models continue to absorb knowledge as their capacity increases. One-shot and few-shot performance make significant gains over zero-shot behavior, matching and exceeding the performance of the SOTA fine-tuned open-domain model, RAG [LPP+20]

and/or the style of their answers are out-of-distribution for GPT-3. Nevertheless, GPT-3 appears able to adapt to this distribution, recovering strong performance in the few-shot setting.

On Natural Questions (NQs) GPT-3 achieves 14.6% in the zero-shot setting, 23.0% in the one-shot setting, and 29.9% in the few-shot setting, compared to 36.6% for fine-tuned T5 11B+SSM. Similar to WebQS, the large gain from zero-shot to few-shot may suggest a distribution shift, and may also explain the less competitive performance compared to TriviaQA and WebQS. In particular, the questions in NQs tend towards very fine-grained knowledge on Wikipedia specifically which could be testing the limits of GPT-3’s capacity and broad pretraining distribution.

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

3.3 Translation

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

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 in a natural way, combining them on a word, sentence, and document level. GPT-3 also uses a single training objective which is not customized or designed for any task in particular. However, our one / few-shot settings aren’t strictly comparable to prior unsupervised work since they make use of a small amount of paired examples (1 or 64). This corresponds to up to a page or two of in-context training data.

Results are shown in Table 3.4. Zero-shot GPT-3, which only receives on a natural language description of the task, still underperforms recent unsupervised NMT results. However, providing only a single example demonstration for

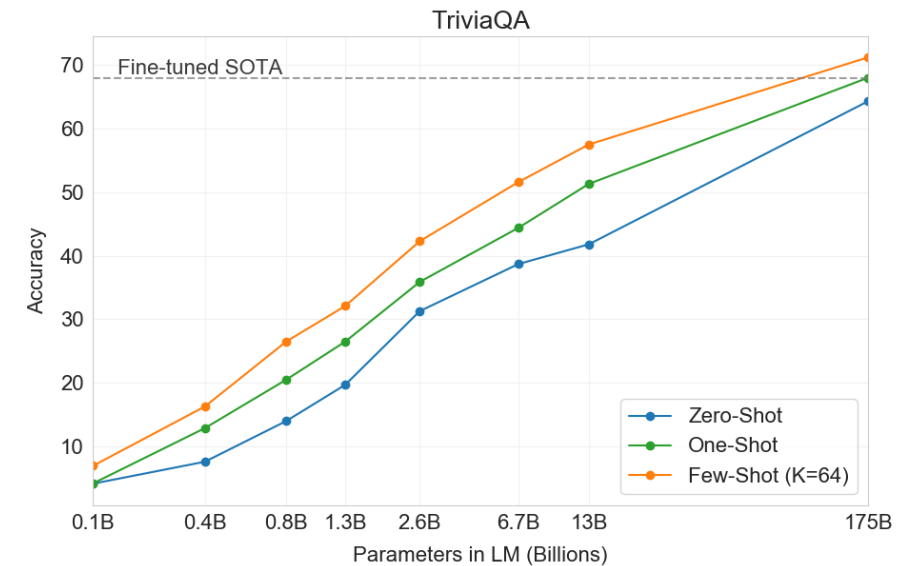


图 3.3: 在 TriviaQA 上，GPT-3 的性能随着模型大小的增加而平稳提升，表明语言模型在容量提升时持续吸收知识。单样本和少样本性能显著优于零样本表现，达到并超越了 SOTA 开放域微调模型 RAG 的性能 [LPP+20]

和/或其答案的风格对于 GPT-3 来说是分布外 (out-of-distribution)。然而，GPT-3 似乎能够适应这种分布，在少样本设置 (few-shot setting) 中恢复出色的性能。

在自然问题 (NQs) 上，GPT-3 在零样本设置中达到 14.6%，单样本设置中达到 23.0%，少样本设置中达到 29.9%，而微调 T5 11B+SSM 则达到 36.6%。与 WebQS 类似，从零样本到少样本的显著提升可能暗示分布偏移，也可能解释了其相较于 TriviaQA 和 WebQS 的竞争力不足。特别是，NQs 中的问题倾向于针对维基百科上非常细粒度的知识，这可能正在测试 GPT-3 的容量和广泛预训练分布的极限。

总体而言，在三个数据集中的其中一个上，GPT-3 的单样本学习表现与开放域微调的 SOTA 相当。在其他两个数据集上，尽管没有使用微调，它的性能也接近闭卷 SOTA 的性能。在所有三个数据集上，我们发现性能随着模型大小的增加非常平稳地提升 (图 3.3 和附录 H，图 H.7)，这可能反映了模型容量直接转化为模型参数中吸收的更多‘知识’这一观点。

3.3 翻译

对于 GPT-2，由于容量限制，我们使用了一个多语言文档集合进行过滤，以生成仅包含英语的数据集。即使经过这样的过滤，GPT-2 也显示出一些多语言能力，并且在翻译法语和英语时表现非同小可，尽管它仅在 10 兆字节的剩余法语文本上进行了训练。由于我们从 GPT-2 到 GPT-3 的容量增加了两个数量级以上，我们也扩展了训练数据集的范围，以包含更多其他语言的代表性内容，尽管这仍然是一个需要进一步改进的领域。如 2.2 节所述，我们的大部分数据是从原始 Common Crawl 中提取的，仅进行了基于质量的过滤。尽管 GPT-3 的训练数据仍然是主要的英语 (按字数计算占 93%)，但它也包括了 7% 的其他语言文本。这些语言在补充材料中有详细说明。为了更好地理解翻译能力，我们还扩展了分析，包括另外两种常见的语言：德语和罗马尼亚语。

现有的无监督机器翻译方法通常结合在单语数据集对上预训练与回译 [SHB15] 来以受控的方式连接这两种语言。相比之下，GPT-3 从混合多种语言的天然方式混合的训练数据中学习，在单词、句子和文档级别上组合它们。GPT-3 还使用一个单一的训练目标，该目标没有针对任何特定任务进行定制或设计。然而，我们的单样本/少样本设置与先前的无监督工作并不严格可比，因为它们使用少量配对示例 (1 或 64)。这对应于最多一页或两页的上下文训练数据。

结果如表 3.4 所示。零样本 GPT-3，它只接收任务的自然语言描述，仍然表现不如最近的无监督 NMT 结果。然而，仅提供一个示例演示

Setting	En→Fr	Fr→En	En→De	De→En	En→Ro	Ro→En
SOTA (Supervised)	45.6^a	35.0 ^b	41.2^c	40.2 ^d	38.5^e	39.9^e
XLM [LC19]	33.4	33.3	26.4	34.3	33.3	31.8
MASS [STQ+19]	<u>37.5</u>	34.9	28.3	35.2	<u>35.2</u>	33.1
mBART [LGG+20]	-	-	<u>29.8</u>	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	<u>39.2</u>	29.7	<u>40.6</u>	21.0	<u>39.5</u>

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 Appendix H. Underline indicates an unsupervised or few-shot SOTA, bold indicates supervised SOTA with relative confidence. ^a[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]

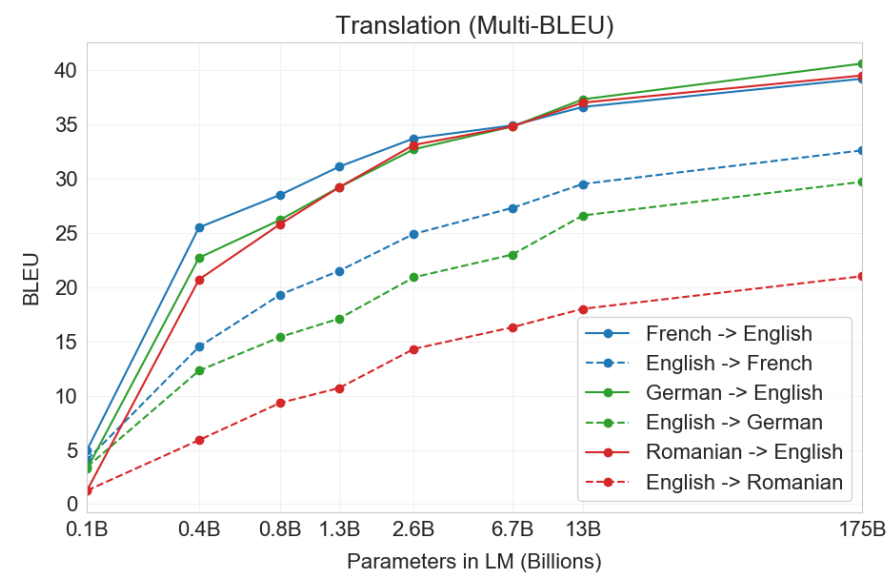


Figure 3.4: Few-shot translation performance on 6 language pairs as model capacity increases. There is a consistent trend of improvement across all datasets as the model scales, and as well as tendency for translation into English to be stronger than translation from English.

设置	En→Fr	Fr→En	En→De	De→En	En→Ro	Ro→En
SOTA (监督)	45.6^a	35.0 ^b	41.2^c	40.2 ^d	38.5^e	39.9^e
XLM [LC19]	33.4	33.3	26.4	34.3	33.3	31.8
MASS [STQ+19]	<u>37.5</u>	34.9	28.3	35.2	<u>35.2</u>	33.1
mBART [LGG+20]	-	-	<u>29.8</u>	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单样本	28.3	33.7	26.2	30.4	20.6	38.6
GPT-3少样本	32.6	<u>39.2</u>	29.7	<u>40.6</u>	21.0	<u>39.5</u>

表3.4: GPT-3在英语翻译上比之前的无监督NMT工作高5个BLEU，反映了它作为英语语言模型的强大能力。我们在WMT’14 Fr↔En、WMT’16 De↔En和WMT’16 Ro↔En数据集上报告BLEU分数，使用multi-bleu.perl测量，并使用XLM的标记化方式，以便与之前的无监督NMT工作最紧密地比较。附录H中报告了SacreBLEU^f [Pos18] 结果。下划线表示无监督或少样本SOTA，粗体表示有监督SOTA（相对置信度）。^a[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]

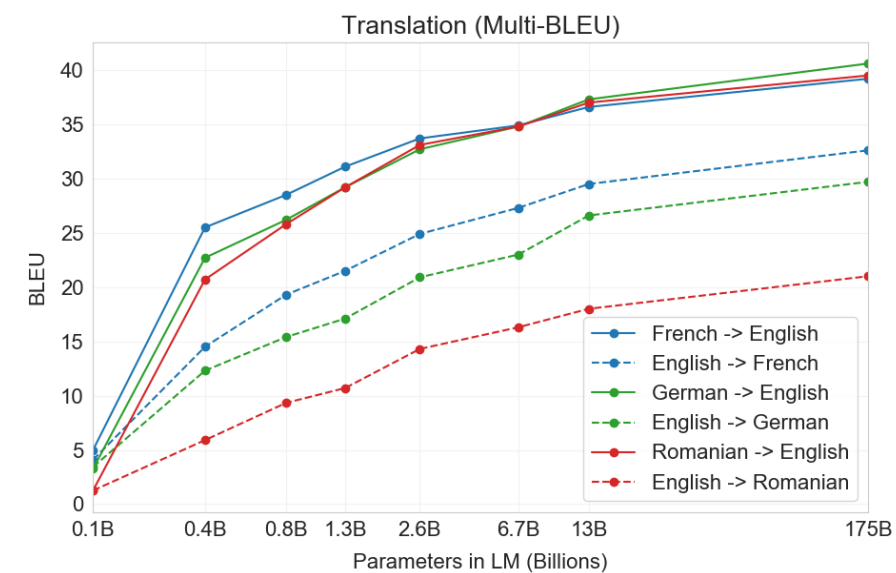


图3.4: 随着模型容量的增加，在6种语言对上的少样本翻译性能。随着模型规模的扩大，所有数据集都呈现出一致的改进趋势，并且英语翻译的倾向性通常强于从英语翻译。

Setting	Winograd	Winogrande (XL)
Fine-tuned SOTA	90.1^a	84.6^b
GPT-3 Zero-Shot	88.3*	70.2
GPT-3 One-Shot	89.7*	73.2
GPT-3 Few-Shot	88.6*	77.7

Table 3.5: Results on the WSC273 version of Winograd schemas and the adversarial Winogrande dataset. See Section 4 for details on potential contamination of the Winograd test set. ^a[SBBC19] ^b[LYN+20]

设置	Winograd	Winogrande (XL)
微调SOTA	90.1^a	84.6^b
GPT-3零样本	88.3*	70.2
GPT-3单样本	89.7*	73.2
GPT-3少样本	88.6*	77.7

表 3.5: 在 Winograd schema 的 WSC273 版本和对抗性 Winogrande 数据集上的结果。参见第 4 节关于 Winograd 测试集潜在污染的信息。 ^a [SBBC19] ^b [LYN+20]



Figure 3.5: Zero-, one-, and few-shot performance on the adversarial Winogrande dataset as model capacity scales. Scaling is relatively smooth with the gains to few-shot learning increasing with model size, and few-shot GPT-3 175B is competitive with a fine-tuned RoBERTa-large.

each translation task improves performance by over 7 BLEU and nears competitive performance with prior work. GPT-3 in the full few-shot setting further improves another 4 BLEU resulting in similar average performance to prior unsupervised NMT work. GPT-3 has a noticeable skew in its performance depending on language direction. For the three input languages studied, GPT-3 significantly outperforms prior unsupervised NMT work when translating into English but underperforms when translating in the other direction. Performance on En-Ro is a noticeable outlier at over 10 BLEU worse than prior unsupervised NMT work. This could be a weakness due to reusing the byte-level BPE tokenizer of GPT-2 which was developed for an almost entirely English training dataset. For both Fr-En and De-En, few shot GPT-3 outperforms the best supervised result we could find but due to our unfamiliarity with the literature and the appearance that these are un-competitive benchmarks we do not suspect those results represent true state of the art. For Ro-En, few shot GPT-3 performs within 0.5 BLEU of the overall SOTA which is achieved by a combination of unsupervised pretraining, supervised finetuning on 608K labeled examples, and backtranslation [LHCG19b].

Finally, across all language pairs and across all three settings (zero-, one-, and few-shot), there is a smooth trend of improvement with model capacity. This is shown in Figure 3.4 in the case of few-shot results, and scaling for all three settings is shown in Appendix H.

3.4 Winograd-Style Tasks

The Winograd Schemas Challenge [LDM12] is a classical task in NLP that involves determining which word a pronoun refers to, when the pronoun is grammatically ambiguous but semantically unambiguous to a human. Recently fine-tuned language models have achieved near-human performance on the original Winograd dataset, but more difficult versions

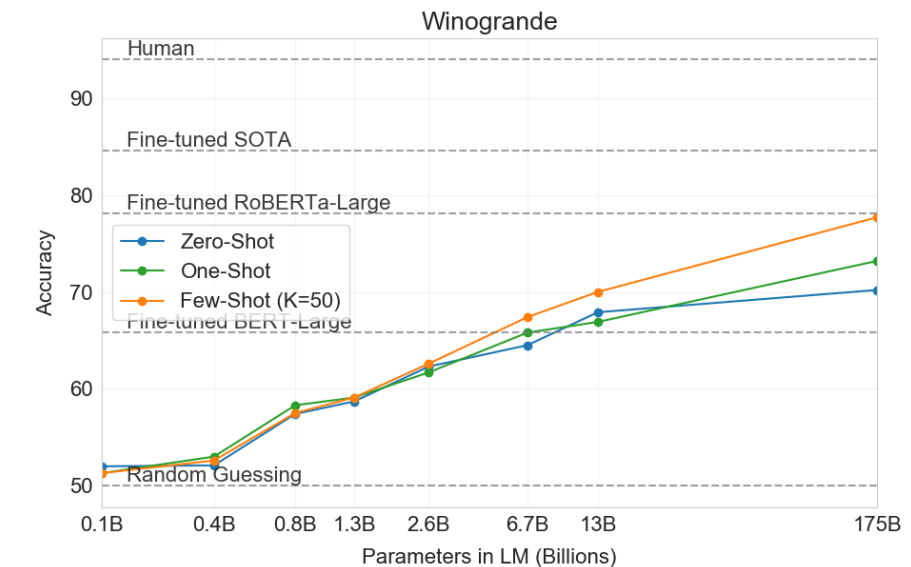


图 3.5: 在对抗性 Winogrande 数据集上，随着模型容量的扩展，零样本、单样本和少样本性能。随着模型容量的增加，少样本学习的收益相对平滑，少样本 GPT-3 175B 与微调的 RoBERTa-Large 具有竞争力。

每个翻译任务通过 7 BLEU 提高了性能，并接近现有工作的竞争性能。在完整的少样本设置中，GPT-3 进一步提高了 4 BLEU，导致其平均性能与现有的无监督 NMT 工作相似。GPT-3 的性能明显取决于语言方向。对于研究的三个输入语言，当翻译成英语时，GPT-3 显著优于现有的无监督 NMT 工作，但在其他方向上表现较差。En-Ro 的性能是一个明显的异常值，比现有的无监督 NMT 工作低 10 BLEU 以上。这可能是由于重用了 GPT-2 的字节级 BPE 分词器，该分词器是为几乎完全是英语的训练数据集开发的弱点。对于 Fr-En 和 De-En，少样本 GPT-3 优于我们找到的最佳监督结果，但由于我们对文献不熟悉，并且这些看起来是非竞争基准测试，所以我们不认为这些结果代表真正的 SOTA。对于 Ro-En，少样本 GPT-3 的性能在 0.5 BLEU 内于整体 SOTA，该 SOTA 是通过无监督预训练、在 608K 标记示例上的监督微调和回译 [LHCG19b]。

最后，在所有语言对和所有三种设置（零样本、单样本和少样本）中，模型容量都呈现出平滑的改进趋势。这在图3.4中对于少样本结果的情况所示，而针对所有三种设置的扩展则显示在附录H。

3.4 Winograd风格任务

Winograd 句式挑战 [LDM12] 是一项经典的 NLP 任务，它涉及确定代词指代哪个词，当代词在语法上模糊但在人类看来语义明确时。最近微调的语言模型在原始 Winograd 数据集上实现了接近人类的性能，但更难的版本

Setting	PIQA	ARC (Easy)	ARC (Challenge)	OpenBookQA
Fine-tuned SOTA	79.4	92.0 [KKS+20]	78.5 [KKS+20]	87.2 [KKS+20]
GPT-3 Zero-Shot	80.5*	68.8	51.4	57.6
GPT-3 One-Shot	80.5*	71.2	53.2	58.8
GPT-3 Few-Shot	82.8*	70.1	51.5	65.4

Table 3.6: GPT-3 results on three commonsense reasoning tasks, PIQA, ARC, and OpenBookQA. GPT-3 Few-Shot PIQA result is evaluated on the test server. See Section 4 for details on potential contamination issues on the PIQA test set.

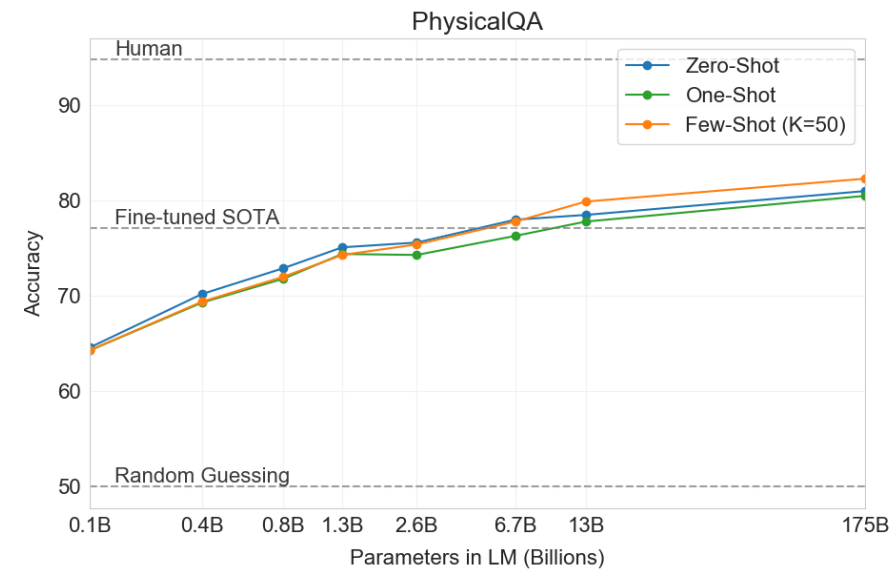


Figure 3.6: GPT-3 results on PIQA in the zero-shot, one-shot, and few-shot settings. The largest model achieves a score on the development set in all three conditions that exceeds the best recorded score on the task.

such as the adversarially-mined Winogrande dataset [SBBC19] still significantly lag human performance. We test GPT-3’s performance on both Winograd and Winogrande, as usual in the zero-, one-, and few-shot setting.

On Winograd we test GPT-3 on the original set of 273 Winograd schemas, using the same “partial evaluation” method described in [RWC+19]. Note that this setting differs slightly from the WSC task in the SuperGLUE benchmark, which is presented as binary classification and requires entity extraction to convert to the form described in this section. On Winograd GPT-3 achieves 88.3%, 89.7%, and 88.6% in the zero-shot, one-shot, and few-shot settings, showing no clear in-context learning but in all cases achieving strong results just a few points below state-of-the-art and estimated human performance. We note that contamination analysis found some Winograd schemas in the training data but this appears to have only a small effect on results (see Section 4).

On the more difficult Winogrande dataset, we do find gains to in-context learning: GPT-3 achieves 70.2% in the zero-shot setting, 73.2% in the one-shot setting, and 77.7% in the few-shot setting. For comparison a fine-tuned RoBERTa model achieves 79%, state-of-the-art is 84.6% achieved with a fine-tuned high capacity model (T5), and human performance on the task as reported by [SBBC19] is 94.0%.

3.5 Common Sense Reasoning

Next we consider three datasets which attempt to capture physical or scientific reasoning, as distinct from sentence completion, reading comprehension, or broad knowledge question answering. The first, PhysicalQA (PIQA) [BZB+19], asks common sense questions about how the physical world works and is intended as a probe of grounded understanding of the world. GPT-3 achieves 81.0% accuracy zero-shot, 80.5% accuracy one-shot, and 82.8% accuracy few-shot (the last measured on PIQA’s test server). This compares favorably to the 79.4% accuracy prior state-of-the-art of a

设置	PIQA	ARC (简单)	ARC (挑战)	OpenBookQA
微调SOTA	79.4	92.0 [KKS+20]	78.5 [KKS+20]	87.2 [KKS+20]
GPT-3零样本	80.5*	68.8	51.4	57.6
GPT-3单样本	80.5*	71.2	53.2	58.8
GPT-3少样本	82.8*	70.1	51.5	65.4

表3.6: GPT-3在三个常识推理任务PIQA、ARC和OpenBookQA上的结果。GPT-3少样本PIQA结果在测试服务器上评估。有关PIQA测试集潜在污染问题的详细信息，请参见第4节。

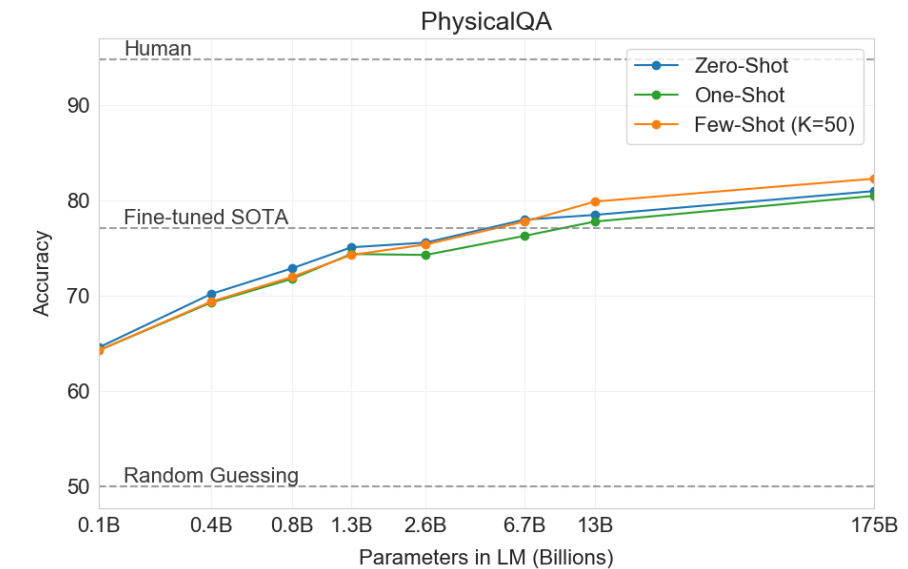


图3.6: GPT-3结果 在零样本、单样本和少样本设置下PIQA的结果。最大模型取得了所有三种条件下的实体集均超过了该任务的最佳记录分数。

例如对抗性挖掘的 Winogrande 数据集 [SBBC19] 仍然显著落后于人类性能。我们像往常一样在零样本、单样本和少样本设置中测试 GPT-3 在 Winograd 和 Winogrande 上的性能。

在 Winograd 上，我们使用原始的 273 个 Winograd 方案测试 GPT-3，采用在 RWC 中描述的“部分评估”方法。请注意，这个设置与 SuperGLUE 基准中的 WSC 任务略有不同，该任务以二元分类的形式呈现，并需要实体提取才能转换为本节中描述的形式。在 Winograd 上，GPT-3 在零样本、单样本和少样本设置中分别达到 88.3%、89.7% 和 88.6%，显示出没有明显的情境学习，但在所有情况下都取得了仅略低于当前最佳水平和估计人类性能的强劲结果。我们注意到，污染分析发现训练数据中存在一些 Winograd 方案，但这似乎只对结果产生很小的影响（见第 9 节）。

在更困难的 Winogrande 数据集上，我们确实发现了情境学习的收益：GPT-3 在零样本设置中达到 70.2%，在单样本设置中达到 73.2%，在少样本设置中达到 77.7%。为了比较，一个微调的 RoBERTa 模型达到 79%，当前最佳水平是 84.6%，由一个微调的高容量模型（T5）实现，而据报道 SBBC 在该任务上的人类性能为 94.0%。

3.5 常识推理

接下来，我们考虑三个试图捕捉物理或科学推理的数据集，这与句子补全、阅读理解或广泛知识问答有所区别。第一个是 PhysicalQA (PIQA) [BZB+19]，它提出了关于物理世界如何运作的常识性问题，旨在测试对世界的扎实理解。GPT-3 零样本学习的准确率为 81.0%，单样本学习的准确率为 80.5%，少样本学习的准确率为 82.8%（最后在 PIQA 的测试服务器上测量）。这与其他先进技术的 79.4% 准确率相比有优势。

Setting	CoQA	DROP	QuAC	SQuADv2	RACE-h	RACE-m
Fine-tuned SOTA	90.7^a	89.1^b	74.4^c	93.0^d	90.0^e	93.1^e
GPT-3 Zero-Shot	81.5	23.6	41.5	59.5	45.5	58.4
GPT-3 One-Shot	84.0	34.3	43.3	65.4	45.9	57.4
GPT-3 Few-Shot	85.0	36.5	44.3	69.8	46.8	58.1

Table 3.7: Results on reading comprehension tasks. All scores are F1 except results for RACE which report accuracy. ^a[JZC⁺19] ^b[JN20] ^c[AI19] ^d[QIA20] ^e[SPP⁺19]

fine-tuned RoBERTa. PIQA shows relatively shallow scaling with model size and is still over 10% worse than human performance, but GPT-3’s few-shot and even zero-shot result outperform the current state-of-the-art. Our analysis flagged PIQA for a potential data contamination issue (despite hidden test labels), and we therefore conservatively mark the result with an asterisk. See Section 4 for details.

ARC [CCE⁺18] is a dataset of multiple-choice questions collected from 3rd to 9th grade science exams. On the “Challenge” version of the dataset which has been filtered to questions which simple statistical or information retrieval methods are unable to correctly answer, GPT-3 achieves 51.4% accuracy in the zero-shot setting, 53.2% in the one-shot setting, and 51.5% in the few-shot setting. This is approaching the performance of a fine-tuned RoBERTa baseline (55.9%) from UnifiedQA [KKS⁺20]. On the “Easy” version of the dataset (questions which either of the mentioned baseline approaches answered correctly), GPT-3 achieves 68.8%, 71.2%, and 70.1% which slightly exceeds a fine-tuned RoBERTa baseline from [KKS⁺20]. However, both of these results are still much worse than the overall SOTAs achieved by the UnifiedQA which exceeds GPT-3’s few-shot results by 27% on the challenge set and 22% on the easy set.

On OpenBookQA [MCKS18], GPT-3 improves significantly from zero to few shot settings but is still over 20 points short of the overall SOTA. GPT-3’s few-shot performance is similar to a fine-tuned BERT Large baseline on the leaderboard.

Overall, in-context learning with GPT-3 shows mixed results on commonsense reasoning tasks, with only small and inconsistent gains observed in the one and few-shot learning settings for both PIQA and ARC, but a significant improvement is observed on OpenBookQA. GPT-3 sets SOTA on the new PIQA dataset in all evaluation settings.

3.6 Reading Comprehension

Next we evaluate GPT-3 on the task of reading comprehension. We use a suite of 5 datasets including abstractive, multiple choice, and span based answer formats in both dialog and single question settings. We observe a wide spread in GPT-3’s performance across these datasets suggestive of varying capability with different answer formats. In general we observe GPT-3 is on par with initial baselines and early results trained using contextual representations on each respective dataset.

GPT-3 performs best (within 3 points of the human baseline) on CoQA [RCM19] a free-form conversational dataset and performs worst (13 F1 below an ELMo baseline) on QuAC [CHI⁺18] a dataset which requires modeling structured dialog acts and answer span selections of teacher-student interactions. On DROP [DWD⁺19], a dataset testing discrete reasoning and numeracy in the context of reading comprehension, GPT-3 in a few-shot setting outperforms the fine-tuned BERT baseline from the original paper but is still well below both human performance and state-of-the-art approaches which augment neural networks with symbolic systems [RLL⁺19]. On SQuAD 2.0 [RJL18], GPT-3 demonstrates its few-shot learning capabilities, improving by almost 10 F1 (to 69.8) compared to a zero-shot setting. This allows it to slightly outperform the best fine-tuned result in the original paper. On RACE [LXL⁺17], a multiple choice dataset of middle school and high school english examinations, GPT-3 performs relatively weakly and is only competitive with the earliest work utilizing contextual representations and is still 45% behind SOTA.

3.7 SuperGLUE

In order to better aggregate results on NLP tasks and compare to popular models such as BERT and RoBERTa in a more systematic way, we also evaluate GPT-3 on a standardized collection of datasets, the SuperGLUE benchmark [WPN⁺19] [WPN⁺19] [CLC⁺19] [DMST19] [RBG11] [KCR⁺18] [ZLL⁺18] [DGM06] [BHDD⁺06] [GMDD07] [BDD⁺09] [PCC18] [PHR⁺18]. GPT-3’s test-set performance on the SuperGLUE dataset is shown in Table 3.8. In the few-shot setting, we used 32 examples for all tasks, sampled randomly from the training set. For all tasks except WSC

设置	CoQA	DROP	QuAC	SQuADv2	RACE-h	RACE-m
微调SOTA	90.7^a	89.1^b	74.4^c	93.0^d	90.0^e	93.1^e
GPT-3零样本	81.5	23.6	41.5	59.5	45.5	58.4
GPT-3单样本	84.0	34.3	43.3	65.4	45.9	57.4
GPT-3少样本	85.0	36.5	44.3	69.8	46.8	58.1

表3.7: 在阅读理解任务上的结果。所有分数均为F1分数，RACE的结果报告准确率。 ^a[JZC⁺19] ^b[JN20] ^c[AI19] ^d[QIA20] ^e[SPP⁺19]

微调的 RoBERTa。PIQA 在模型大小上表现出相对较浅的扩展性，并且仍然比人类性能低 10% 以上，但 GPT-3 的少样本学习甚至零样本学习结果都优于当前其他先进技术。我们的分析发现 PIQA 存在潜在的数据污染问题（尽管测试标签被隐藏），因此我们保守地用星号标记了结果。详情请参见第4节。

ARC [CCE⁺18] 是一个从3年级到9年级科学考试中收集的多项选择题数据集。在数据集的“挑战”版本（已被过滤掉简单统计或信息检索方法无法正确回答的问题）中，GPT-3在零样本设置下达到51.4%的准确率，单样本设置为53.2%，少样本设置为51.5%。这接近了UnifiedQA中微调的RoBERTa基线（55.9%）的性能 [KKS⁺20]。在数据集的“简单”版本（两种基线方法都能正确回答的问题）中，GPT-3达到68.8%、71.2%和70.1%，略微超过了来自 [KKS⁺20]的微调RoBERTa基线。然而，这两种结果仍然远不如 UnifiedQA实现的总体SOTA，在挑战集上比GPT-3的少样本结果高27%，在简单集上高22%。

在OpenBookQA [MCKS18], GPT-3在从零到少样本设置上显著提升，但仍比整体SOTA（挑战）低20多分。GPT-3的少样本性能与排行榜上微调后的BERT Large基线相似。

总体而言，GPT-3的上下文学习在常识推理任务上表现不一，在PIQA和ARC的少样本设置中仅观察到微小且不一致的提升，但在OpenBookQA上显著改善。GPT-3在新PIQA数据集上所有评估设置中都达到了SOTA（挑战）。

3.6 阅读理解

接下来我们在阅读理解任务上评估GPT-3。我们使用包含抽象式、选择题和基于片段的答案格式，涵盖对话和单向两种设置的5个数据集。观察到GPT-3在这些数据集上的表现差异很大，暗示其能力随不同答案格式而变化。通常，GPT-3的表现与各数据集上初始基线及使用上下文表示训练的早期结果相当。

GPT-3 在 CoQA 上表现最佳（比人类基线高出 3 分） [RCM19] 这是一个自由形式的对话数据集，而在 QuAC 上表现最差（比 ELMo 基线低 13 分） [CHI⁺18] 这是一个需要建模结构化对话行为和教师-学生互动答案跨度的数据集。在 DROP [DWD⁺19]，一个测试阅读理解中离散推理和算术能力的测试集，GPT-3 在少样本设置中优于原始论文中微调的 BERT 基线，但仍然远低于人类性能和通过将神经网络与符号系统增强的 SOTA 方法 [RLL⁺19]。在 SQuAD 2.0 [RJL18], GPT-3 展示了其少样本学习能力，与零样本设置相比，F1 提高了近 10 分（达到 69.8），使其略微优于原始论文中的最佳微调结果。在 RACE [LXL⁺17]，一个中学和高中英语考试的选择数据集，GPT-3 表现相对较弱，仅与最早使用上下文表示的工作具有竞争力，并且仍然落后于 SOTA 45%。

3.7 SuperGLUE

为了更好地在自然语言处理任务上聚合结果，并以更系统的方式与BERT和RoBERTa等流行模型进行比较，我们还评估了 GPT-3在一个标准化的数据集集合——SuperGLUE基准测试上的表现[WPN⁺19] [WPN⁺19] [CLC⁺19] [DMST19] [RBG11] [KCR⁺18] [ZLL⁺18] [DGM06] [BHDD⁺06] [GMDD07] [BDD⁺09] [PCC18] [PHR⁺18]。GPT-3在SuperGLUE数据集上的测试集表现如表3.8所示。在少样本设置中，我们为所有任务使用了32个示例，随机从训练集中采样。对于所有任务除WSC

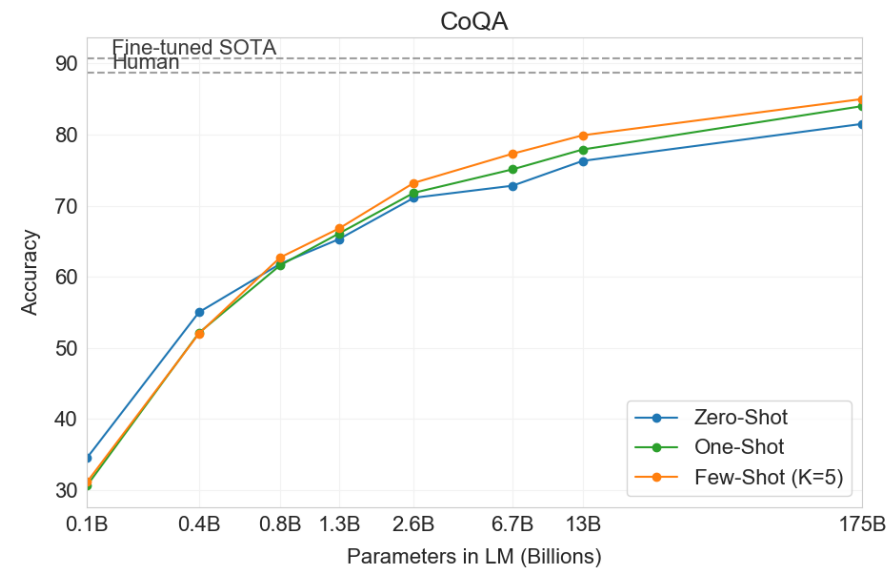


Figure 3.7: GPT-3 results on CoQA reading comprehension task. GPT-3 175B achieves 85 F1 in the few-shot setting, only a few points behind measured human performance and state-of-the-art fine-tuned models. Zero-shot and one-shot performance is a few points behind, with the gains to few-shot being largest for bigger models.

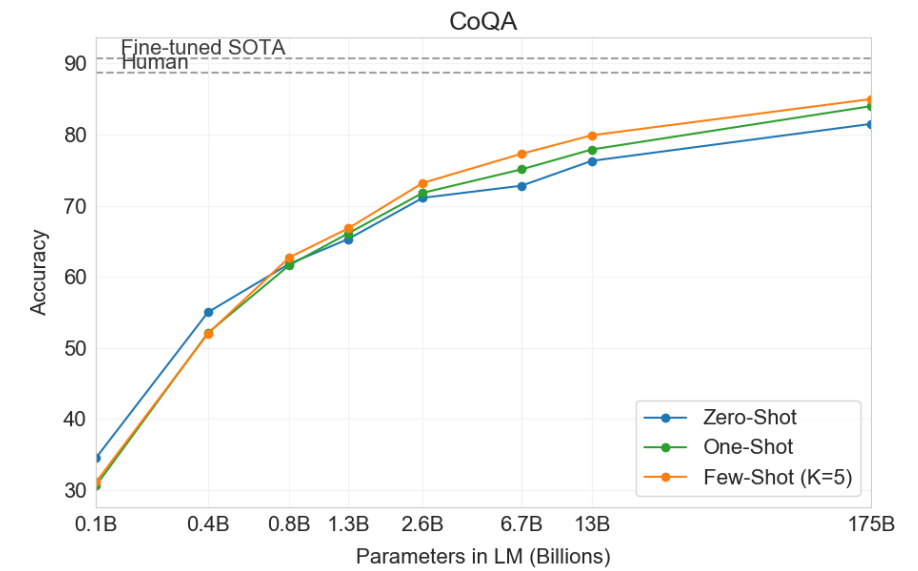


图3.7: GPT-3在CoQA阅读理解任务上的结果。GPT-3 175B在少样本设置中达到85 F1分数，仅比人类性能和最先进的微调模型低几个百分点。零样本学习和单样本学习的性能略低，其中模型越大，从单样本学习转换到少样本学习的收益越大。

	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

	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

Table 3.8: 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.

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

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

表3.8: GPT-3在SuperGLUE上的性能表现与微调基线和SOTA的对比。所有结果均在测试集上报告。GPT-3少样本在每个任务上下文中获得总共32个示例，且不进行梯度更新。

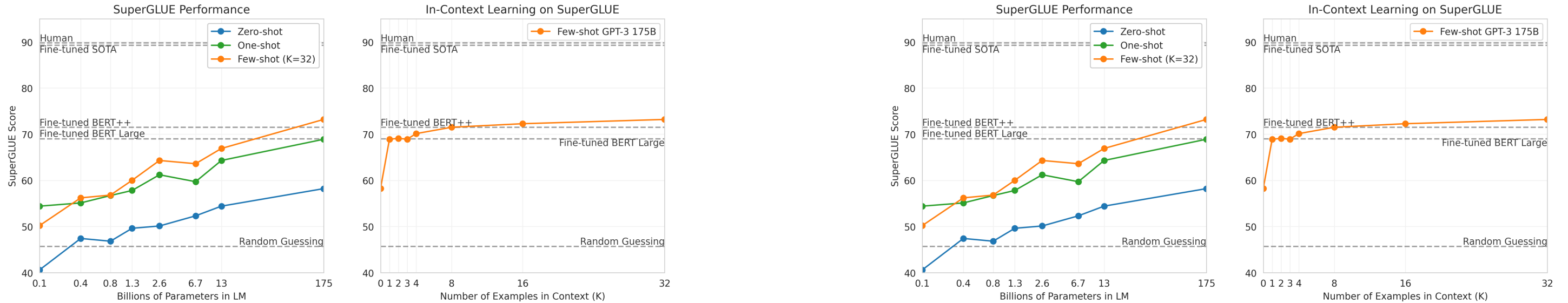


Figure 3.8: Performance on SuperGLUE increases with model size and number of examples in context. A value of $K = 32$ means that our model was shown 32 examples per task, for 256 examples total divided across the 8 tasks in SuperGLUE. We report GPT-3 values on the dev set, so our numbers are not directly comparable to the dotted reference lines (our test set results are in Table 3.8). The BERT-Large reference model was fine-tuned on the SuperGLUE training set (125K 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 630K fine-tuning examples). 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.

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

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

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

3.8 NLI

Natural Language Inference (NLI) [Fyo00] concerns the ability to understand the relationship between two sentences. In practice, this task is usually structured as a two or three class classification problem where the model classifies

图 3.8: 在 SuperGLUE 上的性能随着模型大小和上下文中的示例数量的增加而提高。 一个值 $K = 32$ 表示我们的模型每个任务展示了 32 个示例，总共 256 个示例分布在 SuperGLUE 的 8 个任务中。我们在开发集上报告 GPT-3 的值，因此我们的数字不能直接与虚线参考线进行比较（我们的测试集结果在表 3.8）。BERT-Large 参考模型在 SuperGLUE 训练集（125K 个示例）上进行了微调，而 BERT++ 最初在 MultiNLI（392K 个示例）和 SWAG（113K 个示例）上进行微调，然后再在 SuperGLUE 训练集上进行进一步微调（总共 630K 个微调示例）。我们发现 BERT-Large 和 BERT++ 之间的性能差异大致相当于 GPT-3 每个上下文一个示例与八个示例之间的差异。

并且对于 MultiRC，我们为每个问题在上下文中采样了一个新的示例集来使用。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。

我们观察到 GPT-3 在不同任务上的表现差异很大。在 COPA 和 ReCoRD 上，GPT-3 在单样本学习和少样本设置中实现了接近 SOTA 的性能，其中 COPA 仅落后几个点，并在排行榜上排名第二，而第一名是由一个微调的 110 亿参数模型（T5）保持的。在 WSC 上，性能仍然相对强劲，在少样本设置中达到了 80.1%（请注意，根据第 3.4 节所述，GPT-3 在原始 Winograd 数据集上达到了 88.6%）。在 BoolQ、MultiRC 和 RTE 上，性能合理，大致匹配微调 BERT-Large 的水平。在 CB 上，我们在少样本设置中看到了 75.6% 的迹象。

WiC 是一个显著的弱点，其少量样本性能为 49.4%（随机概率水平）。我们尝试了多种不同的措辞和表述方式来处理 WiC（该任务涉及判断一个词在两个句子中是否具有相同含义），但均未能取得优异性能。这暗示了一种现象，在下一节（讨论 ANLI 基准测试）中将更加清晰——GPT-3 在某些涉及比较两个句子或片段的任务中，特别是在少量样本或单样本设置下表现较弱，例如判断一个词在两个句子中是否以相同方式使用（WiC）、判断一个句子是否是另一个句子的释义，或判断一个句子是否蕴含另一个句子。这也可能解释了 RTE 和 CB 的相对较低分数，这两个任务也采用类似格式。尽管存在这些弱点，GPT-3 仍然在八项任务中的四项上优于微调的 BERT-Large，并且在两项任务上接近由微调的 110 亿参数模型保持的当前最佳水平。

最后，我们注意到少样本 SuperGLUE 分数随着模型大小和上下文中的示例数量的增加而稳步提高，显示出上下文学习的收益不断增加（图 3.8）。我们扩展 K 到每个任务 32 个示例，之后额外的示例将无法可靠地放入我们的上下文中。当扫描 K 的值时，我们发现 GPT-3 每个任务需要少于八个总示例才能在整体 SuperGLUE 分数上优于微调的 BERT-Large。

3.8 NLI

自然语言推理 (NLI) [Fyo00] 关注理解两个句子之间关系的能力。在实际应用中，这项任务通常被结构化为一个二类或三类分类问题，模型需要分类

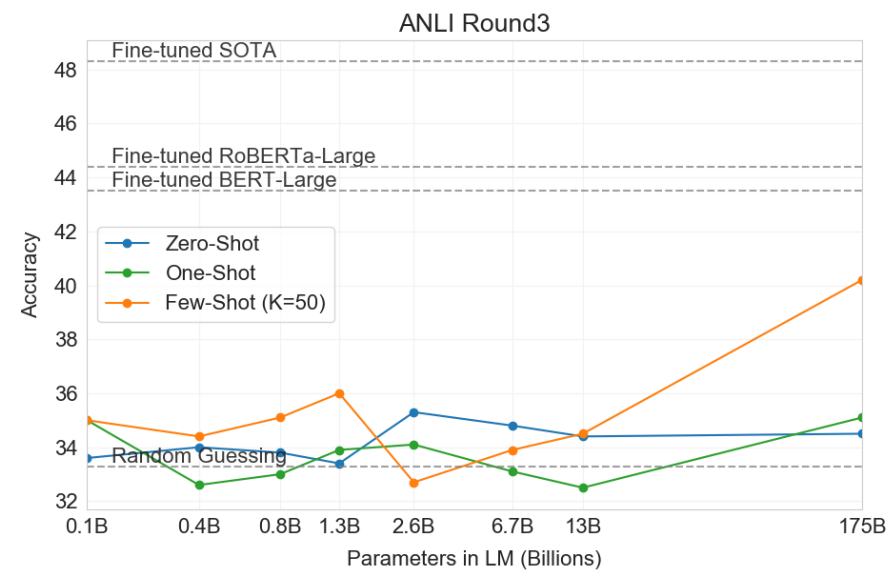


Figure 3.9: Performance of GPT-3 on ANLI Round 3. Results are on the dev-set, which has only 1500 examples and therefore has high variance (we estimate a standard deviation of 1.2%). We find that smaller models hover around random chance, while few-shot GPT-3 175B closes almost half the gap from random chance to SOTA. Results for ANLI rounds 1 and 2 are shown in the appendix.

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

3.9 Synthetic and Qualitative Tasks

One way to probe GPT-3’s range of abilities in the few-shot (or zero- and one-shot) setting is to give it tasks which require it to perform simple on-the-fly computational reasoning, recognize a novel pattern that is unlikely to have occurred in training, or adapt quickly to an unusual task. We devise several tasks to test this class of abilities. First, we test GPT-3’s ability to perform arithmetic. Second, we create several tasks that involve rearranging or unscrambling the letters in a word, tasks which are unlikely to have been exactly seen during training. Third, we test GPT-3’s ability to solve SAT-style analogy problems few-shot. Finally, we test GPT-3 on several qualitative tasks, including using new words in a sentence, correcting English grammar, and news article generation. We will release the synthetic datasets with the hope of stimulating further study of test-time behavior of language models.

3.9.1 Arithmetic

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

- **2 digit addition (2D+)** – The model is asked to add two integers sampled uniformly from [0, 100), phrased in the form of a question, e.g. “Q: What is 48 plus 76? A: 124.”
- **2 digit subtraction (2D-)** – The model is asked to subtract two integers sampled uniformly from [0, 100); the answer may be negative. Example: “Q: What is 34 minus 53? A: -19”.
- **3 digit addition (3D+)** – Same as 2 digit addition, except numbers are uniformly sampled from [0, 1000).

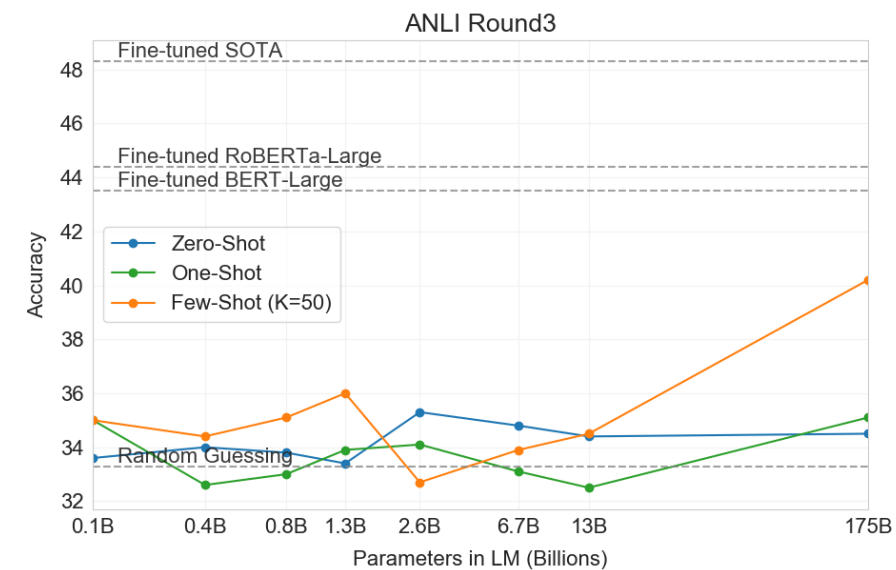


图 3.9: GPT-3 在 ANLI 第三轮上的表现。 结果基于开发集，该集仅有 1500 个示例，因此具有很高的方差（我们估计标准差为 1.2%）。我们发现较小模型的表现接近随机概率，而少样本 GPT-3 175B 几乎缩小了从随机概率到 SOTA 的一半差距。ANLI 第一轮和第二轮的结果在附录中展示。

第二个句子是否在逻辑上遵循第一个句子、与第一个句子矛盾，或可能是真实的（中立）。SuperGLUE 包含一个 NLI 数据集，RTE，用于评估该任务的二分类版本。在 RTE 上，只有 GPT-3 的最大版本在任何评估设置下表现明显优于随机（56%），但在少样本设置中，GPT-3 的表现与单任务微调的 BERT Large 类似。我们还评估了最近引入的对抗性自然语言推理 (ANLI) 数据集 [NWD⁺19]。ANLI 是一个困难的数据集，采用一系列对抗性挖掘的自然语言推理问题，分为三轮 (R1、R2 和 R3)。类似于 RTE，我们所有小于 GPT-3 的模型在 ANLI 上的表现几乎完全是随机概率，即使在少样本设置中 (~33%)，而 GPT-3 本身在第三轮显示出一些进展迹象。ANLI R3 的结果在图 3.9 中突出显示，所有轮次的全结果可以在附录 H 中找到。这些在 RTE 和 ANLI 上的结果表明 NLI 仍然是语言模型的一个非常困难的任务，它们才刚刚开始显示出一些进展迹象。

3.9 合成与定性任务

一种在少样本（或零样本和单样本）设置下探测 GPT-3 能力范围的方法是给它布置需要即时进行简单计算推理、识别训练中不太可能出现的全新模式或快速适应不寻常任务的任务。我们设计了几项任务来测试这类能力。首先，我们测试 GPT-3 进行算术的能力。其次，我们创建了几个涉及重新排列或重组单词字母的任务，这些任务在训练中不太可能被完全遇到。第三，我们测试 GPT-3 在少样本下解决 SAT 风格类比问题的能力。最后，我们在几个定性任务上测试 GPT-3，包括在句子中使用新词、纠正英语语法和新闻文章生成。我们将发布这些合成数据集，希望能促进对语言模型测试时行为的进一步研究。

3.9.1 算术

为了测试 GPT-3 在没有特定任务训练的情况下执行简单算术运算的能力，我们开发了一套包含 10 项测试的小型测试组，这些测试要求 GPT-3 用自然语言提出一个简单的算术问题：

- **2 数字加法 (2D+)** – 模型被要求对从 [0, 100 均匀采样的两个整数进行加法运算，并以问题的形式提出，例如 “Q: 48 加 76 等于多少? A: 124。”
- **2 数字减法 (2D-)** – 模型被要求对从 [0, 100 均匀采样的两个整数进行减法运算；答案可以是负数。示例: “Q: 34 减 53 等于多少? A: -19”。
- **3 两位数加法 (3D+)** – 与两位数加法相同，只是数字从 [0, 1000 中均匀采样。

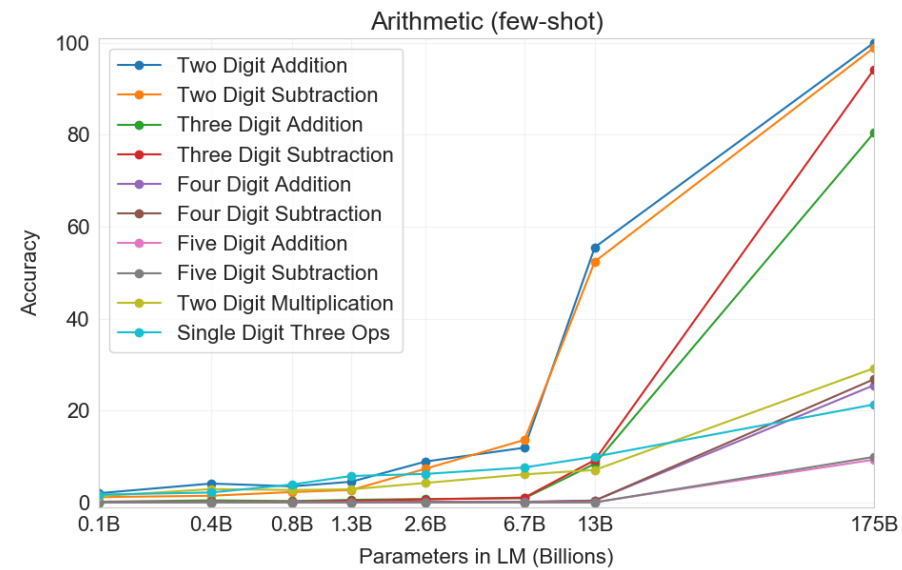


Figure 3.10: Results on all 10 arithmetic tasks in the few-shot settings for models of different sizes. There is a significant jump from the second largest model (GPT-3 13B) to the largest model (GPT-3 175), with the latter being able to reliably accurate 2 digit arithmetic, usually accurate 3 digit arithmetic, and correct answers a significant fraction of the time on 4-5 digit arithmetic, 2 digit multiplication, and compound operations. Results for one-shot and zero-shot are shown in the appendix.

- **3 digit subtraction (3D-)** – Same as 2 digit subtraction, except numbers are uniformly sampled from $[0, 1000)$.
- **4 digit addition (4D+)** – Same as 3 digit addition, except uniformly sampled from $[0, 10000)$.
- **4 digit subtraction (4D-)** – Same as 3 digit subtraction, except uniformly sampled from $[0, 10000)$.
- **5 digit addition (5D+)** – Same as 3 digit addition, except uniformly sampled from $[0, 100000)$.
- **5 digit subtraction (5D-)** – Same as 3 digit subtraction, except uniformly sampled from $[0, 100000)$.
- **2 digit multiplication (2Dx)** – The model is asked to multiply two integers sampled uniformly from $[0, 100)$, e.g. “Q: What is 24 times 42? A: 1008”.
- **One-digit composite (1DC)** – The model is asked to perform a composite operation on three 1 digit numbers, with parentheses around the last two. For example, “Q: What is $6+(4*8)$? A: 38”. The three 1 digit numbers are selected uniformly on $[0, 10)$ and the operations are selected uniformly from $\{+,-,*\}$.

In all 10 tasks the model must generate the correct answer exactly. For each task we generate a dataset of 2,000 random instances of the task and evaluate all models on those instances.

First we evaluate GPT-3 in the few-shot setting, for which results are shown in Figure 3.10. On addition and subtraction, GPT-3 displays strong proficiency when the number of digits is small, achieving 100% accuracy on 2 digit addition, 98.9% at 2 digit subtraction, 80.2% at 3 digit addition, and 94.2% at 3-digit subtraction. Performance decreases as the number of digits increases, but GPT-3 still achieves 25-26% accuracy on four digit operations and 9-10% accuracy on five digit operations, suggesting at least some capacity to generalize to larger numbers of digits. GPT-3 also achieves 29.2% accuracy at 2 digit multiplication, an especially computationally intensive operation. Finally, GPT-3 achieves 21.3% accuracy at single digit combined operations (for example, $9*(7+5)$), suggesting that it has some robustness beyond just single operations.

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

One-shot and zero-shot performance are somewhat degraded relative to few-shot performance, suggesting that adaptation to the task (or at the very least recognition of the task) is important to performing these computations correctly. Nevertheless, one-shot performance is still quite strong, and even zero-shot performance of the full GPT-3 significantly

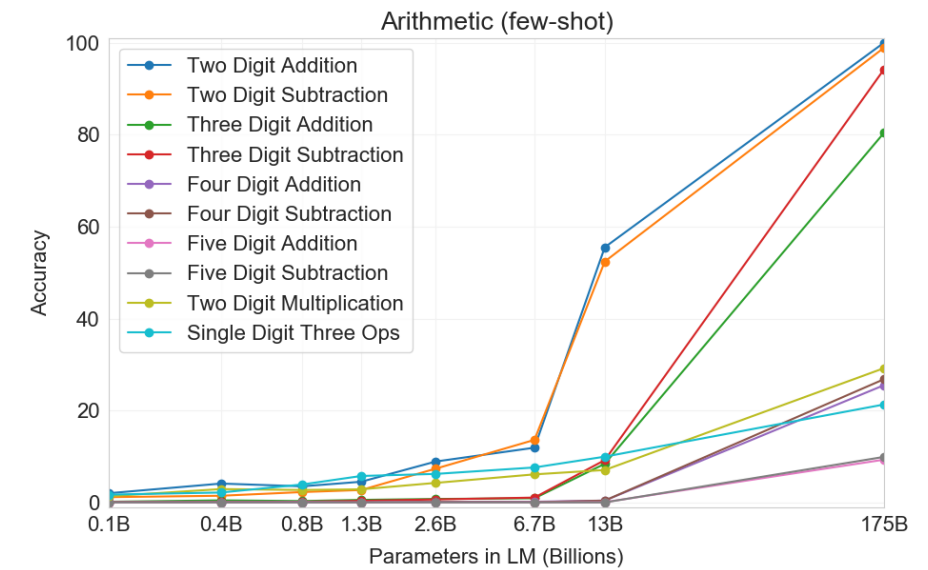


图3.10: 在少样本设置下，不同规模的模型在所有10个算术任务上的结果。从第二大的模型（GPT-3 130亿）到最大的模型（GPT-3 175）之间存在显著跳跃，后者能够可靠地正确2位数字算术，通常正确3位数字算术，并且在4-5位数字算术、2位数字乘法和复合运算中正确答案占显著比例。单样本学习和零样本学习的结果在附录中展示。

- **3 两位数减法 (3D-)** – 与两位数减法相同，只是数字从 $[0, 1000)$ 中均匀采样。
- **4 数字加法 (4D+)** – 与3位数加法相同，只是均匀地从 $[0, 10000)$ 中采样。
- **4 数字减法 (4D-)** – 与3位减法相同，但均匀采样于 $[0, 10000)$ 。
- **5 数字加法 (5D+)** – 与3位数加法相同，只是从 $[0, 100000)$ 中均匀采样。
- **5 数字减法 (5D-)** – 与3位减法相同，但均匀采样于 $[0, 100000)$ 。
- **2 数字乘法 (2Dx)** – 模型被要求对从 $[0, 100)$ 均匀采样的两个整数进行乘法，例如：“Q: 24 乘以 42 等于多少? A: 1008”。
- **一位复合数 (1DC)** – 模型被要求对三个一位数执行复合运算，最后两个数用括号括起来。例如，“Q: $6+(4*8)$ 是什么? A: 38”。这三个一位数在 $[0, 10)$ 上均匀选择，运算从 $\{+,-,*\}$ 中均匀选择。

在所有 10 项任务中，模型必须精确生成正确答案。对于每项任务，我们生成一个包含 2,000 个随机实例的任务数据集，并在这些实例上评估所有模型。

首先，我们在少样本设置中评估 GPT-3，结果如图3.10所示。在加法和减法方面，当数字位数较少时，GPT-3 表现出很强的能力，在两位数加法上达到 100% 准确率，在两位数减法上达到 98.9% 准确率，在三位数加法上达到 80.2%，在三位数减法上达到 94.2%。随着数字位数的增加，性能下降，但 GPT-3 在四位数运算上仍达到 25-26% 的准确率，在五位数运算上达到 9-10% 的准确率，这表明它至少具有一定的泛化到更大位数的能力。GPT-3 在两位数乘法上达到 29.2% 准确率，这是一个特别计算密集型的操作。最后，GPT-3 在单数字组合运算（例如 $9*$ ）上达到 21.3% 准确率，这表明它除了单一操作之外，还有一些鲁棒性。如图 3.10 所示，小型模型在这些任务上都表现不佳——即便是130亿参数模型（仅次于175亿参数完整 GPT-3 的模型）也仅能完成两位数加减法的50%，其他所有运算的成功率都低于10%。

单样本和零样本学习性能相对于少样本学习性能有所下降，这表明任务适应（或者至少是任务识别）对于正确执行这些计算非常重要。尽管如此，单样本学习性能仍然相当强，而且完整GPT-3的零样本学习性能也显著

Setting	2D+	2D-	3D+	3D-	4D+	4D-	5D+	5D-	2Dx	1DC
GPT-3 Zero-shot	76.9	58.0	34.2	48.3	4.0	7.5	0.7	0.8	19.8	9.8
GPT-3 One-shot	99.6	86.4	65.5	78.7	14.0	14.0	3.5	3.8	27.4	14.3
GPT-3 Few-shot	100.0	98.9	80.4	94.2	25.5	26.8	9.3	9.9	29.2	21.3

Table 3.9: Results on basic arithmetic tasks for GPT-3 175B. $\{2,3,4,5\}D\{+,-\}$ is 2, 3, 4, and 5 digit addition or subtraction, 2Dx is 2 digit multiplication. 1DC is 1 digit composite operations. Results become progressively stronger moving from the zero-shot to one-shot to few-shot setting, but even the zero-shot shows significant arithmetic abilities.

Setting	CL	A1	A2	RI	RW
GPT-3 Zero-shot	3.66	2.28	8.91	8.26	0.09
GPT-3 One-shot	21.7	8.62	25.9	45.4	0.48
GPT-3 Few-shot	37.9	15.1	39.7	67.2	0.44

Table 3.10: GPT-3 175B performance on various word unscrambling and word manipulation tasks, in zero-, one-, and few-shot settings. CL is “cycle letters in word”, A1 is anagrams of but the first and last letters, A2 is anagrams of all but the first and last two letters, RI is “Random insertion in word”, RW is “reversed words”.

outperforms few-shot learning for all smaller models. All three settings for the full GPT-3 are shown in Table 3.9, and model capacity scaling for all three settings is shown in Appendix H.

To spot-check whether the model is simply memorizing specific arithmetic problems, we took the 3-digit arithmetic problems in our test set and searched for them in our training data in both the forms " $\langle \text{NUM1} \rangle + \langle \text{NUM2} \rangle =$ " and " $\langle \text{NUM1} \rangle \text{ plus } \langle \text{NUM2} \rangle$ ". Out of 2,000 addition problems we found only 17 matches (0.8%) and out of 2,000 subtraction problems we found only 2 matches (0.1%), suggesting that only a trivial fraction of the correct answers could have been memorized. In addition, inspection of incorrect answers reveals that the model often makes mistakes such as not carrying a “1”, suggesting it is actually attempting to perform the relevant computation rather than memorizing a table.

Overall, GPT-3 displays reasonable proficiency at moderately complex arithmetic in few-shot, one-shot, and even zero-shot settings.

3.9.2 Word Scrambling and Manipulation Tasks

To test GPT-3’s ability to learn novel symbolic manipulations from a few examples, we designed a small battery of 5 “character manipulation” tasks. Each task involves giving the model a word distorted by some combination of scrambling, addition, or deletion of characters, and asking it to recover the original word. The 5 tasks are:

- **Cycle letters in word (CL)** – The model is given a word with its letters cycled, then the “=” symbol, and is expected to generate the original word. For example, it might be given “lyinevitab” and should output “inevitably”.
- **Anagrams of all but first and last characters (A1)** – The model is given a word where every letter except the first and last have been scrambled randomly, and must output the original word. Example: criroptuon = corruption.
- **Anagrams of all but first and last 2 characters (A2)** – The model is given a word where every letter except the first 2 and last 2 have been scrambled randomly, and must recover the original word. Example: opoepnnt → opponent.
- **Random insertion in word (RI)** – A random punctuation or space character is inserted between each letter of a word, and the model must output the original word. Example: s.u!c/c!e.s s i/o/n = succession.
- **Reversed words (RW)** – The model is given a word spelled backwards, and must output the original word. Example: stcejbo → objects.

For each task we generate 10,000 examples, which we chose to be the top 10,000 most frequent words as measured by [Nor09] of length more than 4 characters and less than 15 characters. The few-shot results are shown in Figure 3.11. Task performance tends to grow smoothly with model size, with the full GPT-3 model achieving 66.9% on removing

设置	2D+	2D-	3D+	3D-	4D+	4D-	5D+	5D-	2Dx	1DC
GPT-3零样本	76.9	58.0	34.2	48.3	4.0	7.5	0.7	0.8	19.8	9.8
GPT-3单样本	99.6	86.4	65.5	78.7	14.0	14.0	3.5	3.8	27.4	14.3
GPT-3少样本	100.0	98.9	80.4	94.2	25.5	26.8	9.3	9.9	29.2	21.3

表3.9: GPT-3 175B在基本算术任务上的结果。 $\{2,3,4,5\}D\{+,-\}$ 是2位、3位、4位和5位加减法，2Dx是2位乘法。1DC是1位复合运算。结果从零样本学习到单样本学习到少样本设置逐渐增强，但即使是零样本学习也显示出显著的算术能力。

设置	CL	A1	A2	RI	RW
GPT-3零样本	3.66	2.28	8.91	8.26	0.09
GPT-3单样本	21.7	8.62	25.9	45.4	0.48
GPT-3少样本	37.9	15.1	39.7	67.2	0.44

表 3.10: GPT-3 175B 在各种字母重组和单词操作任务上的表现，在零样本、单样本和少样本设置下。CL是“循环单词中的字母”，A1是除了首尾字母之外的其他字母的字母重组，A2是除了首尾两个字母之外的所有字母的字母重组，RI是“单词中随机插入”，RW是“反转单词”。

优于所有较小模型的少样本学习。完整GPT-3的所有三种设置都显示在表3.9中，三种设置中的模型容量缩放显示在附录H中。

为了检查模型是否只是在记忆特定的算术问题，我们取了测试集中的3位数字算术问题，并在训练数据中搜索了它们，形式为“ $\langle \text{NUM1} \rangle + \langle \text{NUM2} \rangle =$ ”和“ $\langle \text{NUM1} \rangle \text{ plus } \langle \text{NUM2} \rangle$ ”。在2000个加法问题中，我们只找到了17个匹配（0.8%），在2000个减法问题中，我们只找到了2个匹配（0.1%），这表明只有极小一部分的正确答案可能被记忆了。此外，检查错误答案表明，模型经常犯诸如不进位“1”的错误，这表明它实际上是在尝试执行相关计算，而不是记忆一个表。

总体而言，GPT-3在少样本、单样本甚至零样本设置下，对中等复杂度的算术表现出合理的能力。

3.9.2 单词重组与操作任务

为了测试GPT-3从少量示例中学习新型符号操作的能力，我们设计了一组5个“字符操作”任务。每个任务都涉及向模型提供被某种组合的重组、添加或删除字符所扭曲的单词，并要求它恢复原始单词。这5个任务包括：

- **单词字母循环 (CL)** – 模型被给定一个字母顺序被循环的单词，然后是“=”符号，并预期生成原始单词。例如，它可能被给定“lyinevitab”，应输出“inevitably”。
- **首尾字符外所有字母的字母重组 (A1)** – 模型被给定一个首尾字符以外的所有字母被随机重组的单词，必须输出原始单词。示例：criroptuon =corruption。
- **除首尾两个字符外所有字符的字母重组 (A2)** – 模型被给出一个单词，其中除了首尾两个字符外，其余每个字母都已被随机打乱，并且必须恢复原始单词。示例：opoepnnt → opponent。
- **单词中随机插入 (RI)** – 在单词的每个字母之间随机插入一个标点符号或空格字符，并且模型必须输出原始单词。示例：s.u!c/c!e.s s i/o/n = succession。
- **单词反转 (RW)**– 模型被给出一个反向拼写的单词，并且必须输出原始单词。示例：stcejbo → objects。

对于每个任务，我们生成了10,000个示例，我们选择的是按频率最高的前10,000个单词，这是根据[北09]长度超过4个字符且少于15个字符的。少样本结果如图3.11所示。任务性能通常随着模型大小平滑增长，完整的GPT-3模型在移除

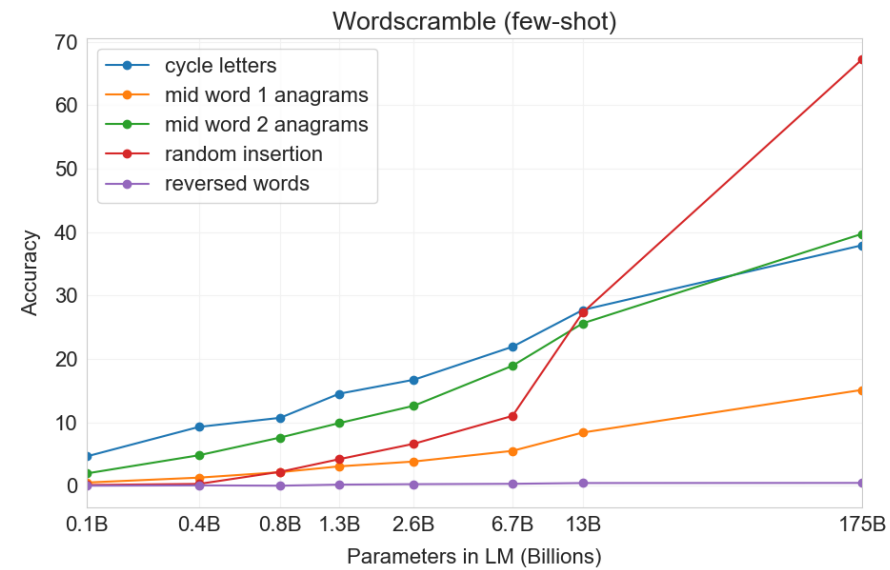


Figure 3.11: Few-shot performance on the five word scrambling tasks for different sizes of model. There is generally smooth improvement with model size although the random insertion task shows an upward slope of improvement with the 175B model solving the task the majority of the time. Scaling of one-shot and zero-shot performance is shown in the appendix. All tasks are done with $K = 100$.

random insertions, 38.6% on cycling letters, 40.2% on the easier anagram task, and 15.1% on the more difficult anagram task (where only the first and last letters are held fixed). None of the models can reverse the letters in a word.

In the one-shot setting, performance is significantly weaker (dropping by half or more), and in the zero-shot setting the model can rarely perform any of the tasks (Table 3.10). This suggests that the model really does appear to learn these tasks at test time, as the model cannot perform them zero-shot and their artificial nature makes them unlikely to appear in the pre-training data (although we cannot confirm this with certainty).

We can further quantify performance by plotting “in-context learning curves”, which show task performance as a function of the number of in-context examples. We show in-context learning curves for the Symbol Insertion task in Figure 1.2. We can see that larger models are able to make increasingly effective use of in-context information, including both task examples and natural language task descriptions.

Finally, it is worth adding that solving these tasks requires character-level manipulations, whereas our BPE encoding operates on significant fractions of a word (on average ~ 0.7 words per token), so from the LM’s perspective succeeding at these tasks involves not just manipulating BPE tokens but understanding and pulling apart their substructure. Also, CL, A1, and A2 are not bijective (that is, the unscrambled word is not a deterministic function of the scrambled word), requiring the model to perform some search to find the correct unscrambling. Thus, the skills involved appear to require non-trivial pattern-matching and computation.

3.9.3 SAT Analogies

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

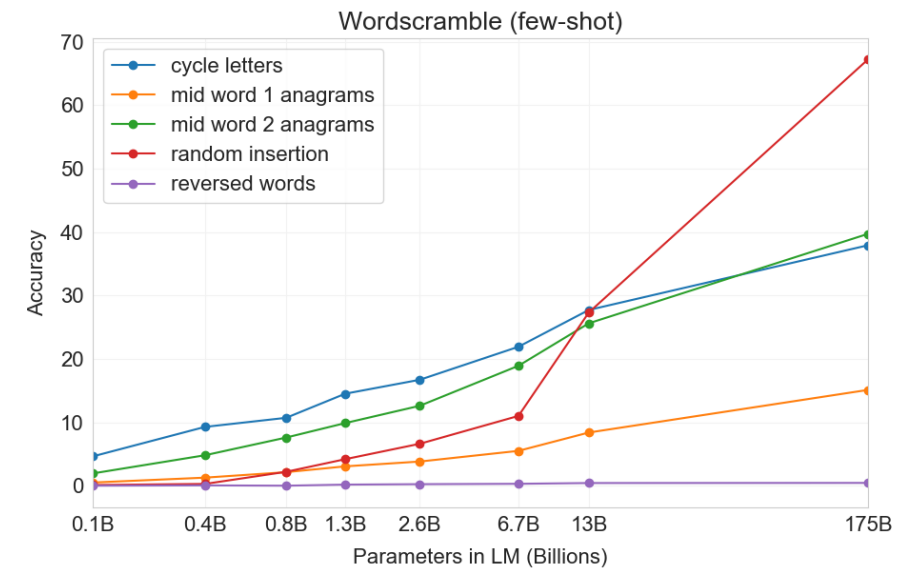


图3.11: 在不同模型大小下，五个单词重组任务的少量样本性能。随着模型大小的增加，性能通常呈现平滑提升，尽管随机插入任务在175B模型上显示出随着模型规模增加而性能提升的上升趋势，该模型多数情况下能够解决该任务。单样本学习和零次学习性能的扩展性在附录中展示。所有任务均使用 $K = 100$ 完成。

随机插入上达到38.6%，在循环字母上达到40.2%，在较容易的字母重组任务上达到40.2%，在较困难的字母重组任务（仅固定首尾字母）上达到15.1%。没有模型能够颠倒单词中的字母。

在单样本设置中，性能显著较弱（降低了一半或更多），而在零样本设置中，模型很少能完成任何任务（表3.10）。这表明模型确实似乎在测试时学习这些任务，因为模型零样本无法完成它们，并且它们的人工性质使它们不太可能在预训练数据中出现（尽管我们无法确定这一点）。

我们可以通过绘制“上下文学习曲线”来进一步量化性能，这些曲线显示了任务性能与上下文示例数量的函数关系。我们在图1.2中展示了符号插入任务的上下文学习曲线。我们可以看到，更大的模型能够越来越有效地利用上下文信息，包括任务示例和自然语言任务描述。

最后，值得补充的是，解决这些任务需要字符级操作，而我们的BPE编码作用于单词的重要部分（平均每个 token 对应 ~ 0.7 个单词），因此从语言模型的角度来看，成功完成这些任务不仅涉及操作 BPE token，还需要理解和拆解它们的子结构。此外，CL、A1和A2不是双射的（也就是说，未打乱的字词不是打乱的字词的确定性函数），需要模型进行一些搜索才能找到正确的解。因此，所涉及的技能似乎需要非平凡的模式匹配和计算。

3.9.3 SAT类比

为了在另一个相对于典型文本分布而言有些不寻常的任务上测试GPT-3，我们收集了一套374个“SAT类比”问题 [TLBS03]。类比是一种选择题类型，在2005年之前曾是SAT大学入学考试的一个部分。一个典型的例子是“audacious是到boldness的关系，如同 (a) sanctimonious是到hypocrisy的关系，(b) anonymous是到identity的关系，(c) remorseful是到misdeed的关系，(d) deleterious是到result的关系，(e) impressionable是到temptation的关系”。学生需要选择哪个五词对与原词对具有相同的关系；在这个例子中答案是“sanctimonious是到hypocrisy的关系”。在这个任务上，GPT-3在少样本设置中达到65.2%，在单样本设置中达到59.1%，在零样本设置中达到53.7%，而大学申请者的平均分数是57% [TL05] (随机猜测的结果是20%)。如图3.12所示，结果随着规模提升而改善，与130亿参数模型相比，完整的1750亿参数模型提升了超过10%

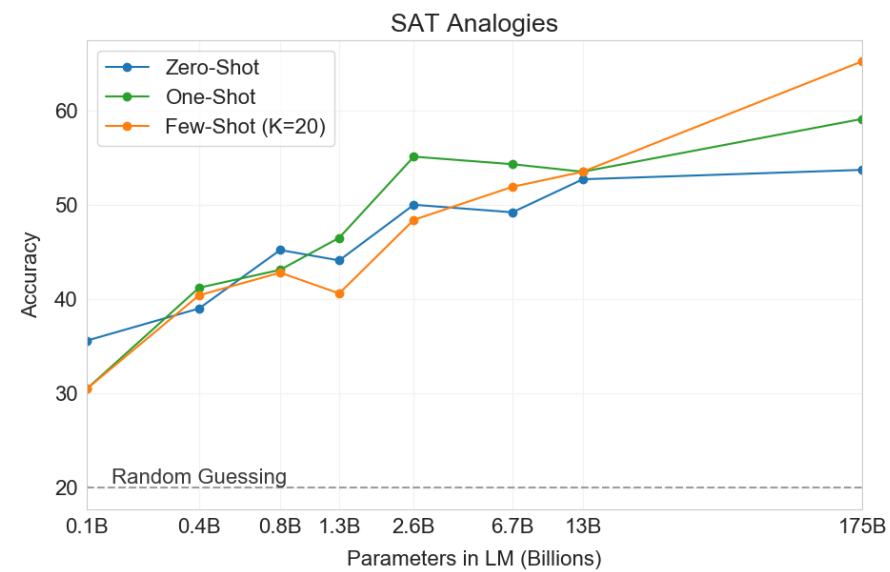


Figure 3.12: Zero-, one-, and few-shot performance on SAT analogy tasks, for different sizes of model. The largest model achieves 65% accuracy in the few-shot setting, and also demonstrates significant gains to in-context learning which are not present in smaller models.

3.9.4 News Article Generation

Previous work on generative language models qualitatively tested their ability to generate synthetic “news articles” by conditional sampling from the model given a human-written prompt consisting of a plausible first sentence for a news story [RWC⁺19]. Relative to [RWC⁺19], the dataset used to train GPT-3 is much less weighted towards news articles, so trying to generate news articles via raw unconditional samples is less effective – for example GPT-3 often interprets the proposed first sentence of a “news article” as a tweet and then posts synthetic responses or follow-up tweets. To solve this problem we employed GPT-3’s few-shot learning abilities by providing three previous news articles in the model’s context to condition it. With the title and subtitle of a proposed next article, the model is able to reliably generate short articles in the “news” genre.

To gauge the quality of news article generation from GPT-3 (which we believe is likely to be correlated with conditional sample generation quality in general), we decided to measure human ability to distinguish GPT-3-generated articles from real ones. Similar work has been carried out by Kreps et al. [KMB20] and Zellers et al. [ZHR⁺19]. Generative language models are trained to match the distribution of content generated by humans, so the (in)ability of humans to distinguish the two is a potentially important measure of quality.³

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 four language models ranging in size from 125M to 175B (GPT-3) parameters (mean length: 200 words). For each model, we presented around 80 US-based participants with a quiz consisting of these real titles and subtitles followed by either the human written article or the article generated by the model⁴. 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”.

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.

³This task is also relevant to the potential misuse of language models discussed in Section 6.1.

⁴We 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 Appendix E for details.

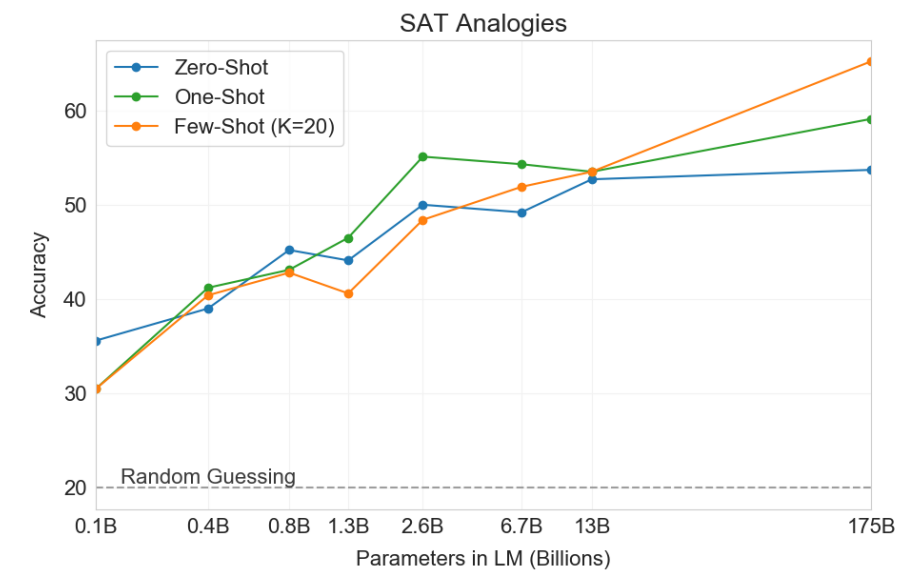


图3.12: SAT类比任务上的零样本、单样本和少样本性能，针对不同规模的模型。最大模型在少样本设置中达到65%准确率，并且也展示了对情境学习的显著提升，而较小模型则没有这种表现。

3.9.4 新闻文章生成

在生成式语言模型方面的先前工作通过从模型中条件采样，给定一个由人类编写的、对新闻故事可能的第一句话组成的提示，定性测试了它们生成合成“新闻文章”的能力 [RWC⁺19]。相对于 [RWC⁺19]，用于训练 GPT-3 的数据集在新闻文章上的权重要小得多，因此通过原始无条件样本生成新闻文章的效果较差——例如 GPT-3 经常将“新闻文章”的建议第一句话解释为推文，然后发布合成响应或后续推文。为了解决这个问题，我们利用了 GPT-3 的少样本学习能力，通过在模型的上下文中提供三个先前的新闻文章来对其进行条件化。通过下一篇文章的标题和副标题，模型能够可靠地生成属于“新闻”类型的短文章。

为了评估 GPT-3 生成的新闻文章的质量（我们相信这与一般条件样本生成质量可能相关），我们决定测量人类区分 GPT-3 生成的文章和真实文章的能力。类似的工作由 Kreps 等人进行过。[KMB20] 以及 Zellers 等人。[ZHR⁺19]。生成式语言模型被训练以匹配人类生成内容的分布，因此人类区分两者的（能）力是一个潜在的重要质量衡量标准。³

为了观察人类检测模型生成文本的能力如何，我们从网站 newser.com (平均长度: 215 词) 中任意选择了 25 篇文章标题和副标题。然后，我们使用四个不同规模的语言模型（参数范围从 125M 到 175B (GPT-3)）生成了这些标题和副标题的补全文本（平均长度: 200 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，测验内容包括这些真实的标题和副标题，随后是人工撰写的文章或模型生成的文章⁴。要求参与者选择文章是“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”或“非常可能由机器撰写”。

我们选定的文章不在模型的训练数据中，且模型输出经过程序格式化和筛选以防止人工挑拣。所有模型都使用相同的上下文来对输出进行条件化，并且都使用相同的上下文大小进行预训练，每个模型的提示都使用了相同的文章标题和副标题。然而，我们还进行了一项实验来控制参与者的努力程度和注意力，该实验采用相同的格式，但涉及故意生成的差模型文章。这是通过从“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。

³这项任务也与第 6.1 节中讨论的语言模型的潜在滥用问题相关。⁴我们想要了解普通网民在检测语言模型输出方面的能力如何，因此我们专注于招募来自美国普通人群的参与者。详情请参见附录 E。

	Mean accuracy	95% Confidence Interval (low, hi)	<i>t</i> compared to control (<i>p</i> -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%

Table 3.11: Human accuracy in identifying whether short (~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).

Mean human accuracy (the ratio of correct assignments to non-neutral assignments per participant) at detecting that the intentionally bad articles were model generated was ~ 86% where 50% is chance level performance. By contrast, mean human accuracy at detecting articles that were produced by the 175B parameter model was barely above chance at ~ 52% (see Table 3.11).⁵ 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.⁶ This is true despite the fact that participants spend more time on each output as model size increases (see Appendix E).

Examples of synthetic articles from GPT-3 are given in Figures 3.14 and 3.15.⁷ Much of the text is—as indicated by 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.

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 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 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.

We found that mean human accuracy at detecting the intentionally bad longer articles from the control model was ~ 88%, while mean human accuracy at detecting the longer articles that were produced by GPT-3 175B was still barely above chance at ~ 52% (see Table 3.12). 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.

3.9.5 Learning and Using Novel Words

A task studied in developmental linguistics [CB78] is the ability to learn and utilize new words, for example using a word in a sentence after seeing it defined only once, or conversely inferring a word’s meaning from only one usage. Here we qualitatively test GPT-3’s ability to do the former. Specifically, we give GPT-3 the definition of a nonexistent word, such as “Gigamuru”, and then ask it to use it in a sentence. We provide one to five previous examples of a (separate)

⁵We 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.

⁶If 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.

⁷Additional non-news samples can be found in Appendix F.

	平均准确率	95%置信度区间 (低, 高)	<i>t</i> 与对照 (<i>p</i> -值)	“我不知道”分配
控制 (故意差模型)	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 27亿	62%	58%–65%	10.4 (5e-19)	7.1%
GPT-3 67亿	60%	56%–63%	11.2 (3e-21)	6.2%
GPT-3 130亿	55%	52%–58%	15.3 (1e-32)	7.1%
GPT-3 175B	52%	49%–54%	16.9 (1e-34)	7.8%

表3.11: 人类识别短 (~200 字) 新闻文章是否为模型生成的准确率。我们发现人类准确率 (通过正确分配与非中立分配的比例衡量) 在控制模型上为86%, 在GPT-3 175B上为52%。该表比较了五个不同模型之间的平均准确率, 并展示了每个模型与控制模型 (一个输出随机性增加的无条件GPT-3小型模型) 之间平均准确率差异的双样本T检验结果。

检测故意生成的差文章是模型生成的平均人类准确率 (正确分配与非中性分配的参与者比例) 为 ~ 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平均人类准确率在 ~ 52% (见表3.11)。⁵ 检测模型生成文本的人类能力似乎随着模型大小的增加而下降: 似乎存在一个趋势, 即模型大小越高, 人类准确率越接近随机水平, 且人类检测GPT-3的能力接近随机水平。⁶ 尽管参与者在模型大小增加时花费在每个输出上的时间更多 (见附录E)。

GPT-3生成的合成文章示例见图3.14和3.15。⁷ 根据评估, 大部分文本——正如所指示的——对人类来说难以与真实的人类内容区分开来。事实性错误可能是文章为模型生成的指标, 因为与人类作者不同, 模型无法访问文章标题所指的具体事实或文章的写作时间。其他指标包括重复、前后不连贯和不寻常的措辞, 尽管这些通常足够微妙以至于不被注意到。

Ippolito 等人关于语言模型检测的相关研究 [IDCBE19]表明, 像 GROVER [ZHR⁺19] 和 GLTR [GSR19] 这样的自动判别器在检测模型生成的文本方面可能比人类评估员更成功。自动检测这些模型可能是未来研究的一个有前景的领域。

Ippolito 等人 [IDCBE19] 还指出, 随着人类观察更多 token, 人类检测模型生成文本的准确率会提高。为了初步调查人类在检测 GPT-3 175B 生成的较长的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 个词, 并使用 GPT-3 生成了这些文章的补全, 平均长度为 498 个词 (比我们最初的实验长 298 个词)。遵循上述方法, 我们进行了两项实验, 每项实验约有 80 名美国参与者, 以比较人类检测 GPT-3 和控制模型生成的文章的能力。

我们发现, 从控制模型中检测出故意编写的更长文章的平均人类准确率是 ~ 88%, 而检测由 GPT-3 175B 生成的更长文章的平均人类准确率仍然略高于随机水平 ~ 52% (参见表3.12)。这表明, 对于大约500字长的新闻文章, GPT-3 仍然会生成人类难以将其与人类撰写的新闻文章区分开的文章。

3.9.5 学习和使用新词

发展语言学中的一个任务 [CB78] 是学习和使用新词的能力, 例如在仅看到一次定义后使用一个词, 或者相反从仅有的一个用法中推断一个词的含义。在这里, 我们定性测试GPT-3做前者的能力。具体来说, 我们给GPT-3一个不存在词的定义, 例如“Gigamuru”, 然后要求它用它造句。我们提供一到五个先前示例 (独立的)

⁵我们使用双样本Student’s T检验来检验每个模型的参与者准确率与控制模型之间的均值是否存在显著差异, 并报告均值差异的标准化结果 (以*t*统计量表示) 和*p*值。⁶如果一个模型持续生成比人类文章更令人印象深刻的内容, 那么在这个任务上的人类性能可能会低于50%。事实上, 许多单个参与者在任务上的得分低于50%。⁷其他非新闻样本可以在附录F中找到。

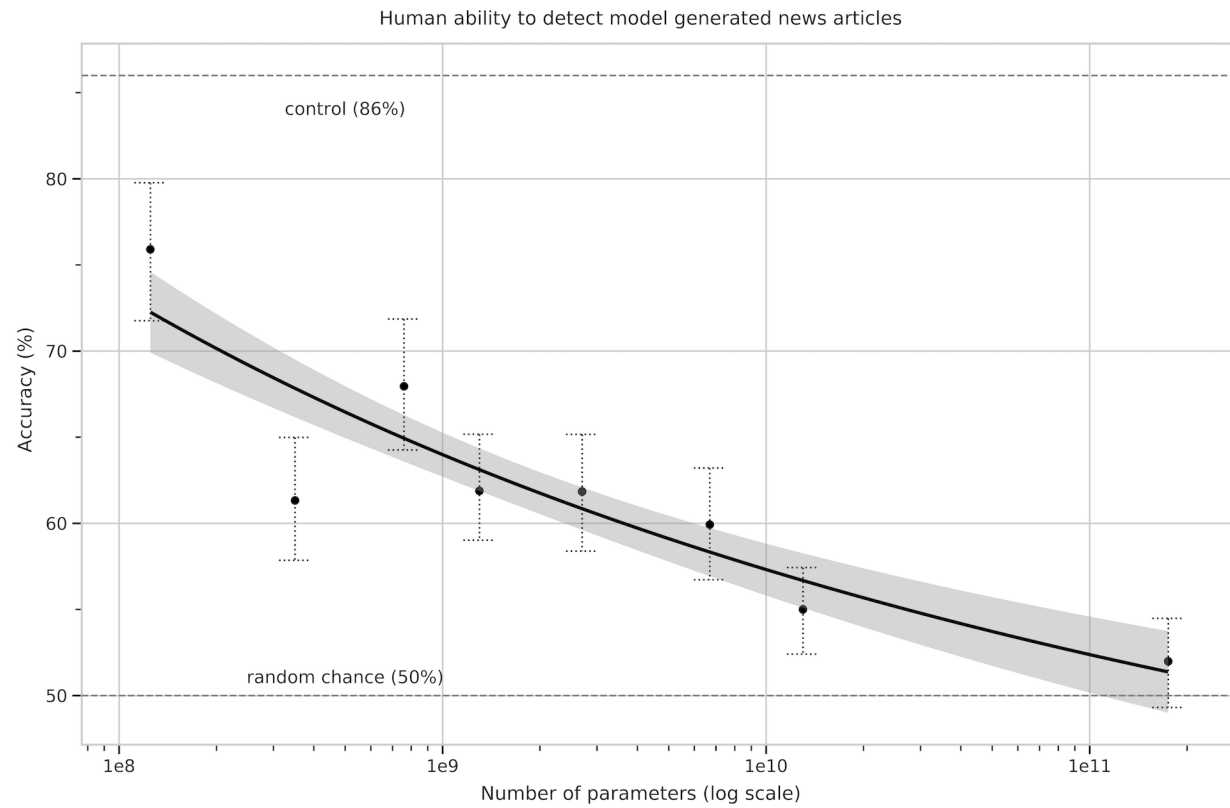


Figure 3.13: People’s ability to identify whether news articles are model-generated (measured by the ratio of correct assignments to non-neutral assignments) decreases as model size increases. Accuracy on the outputs on the deliberately-bad control model (an unconditioned GPT-3 Small model with higher 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.

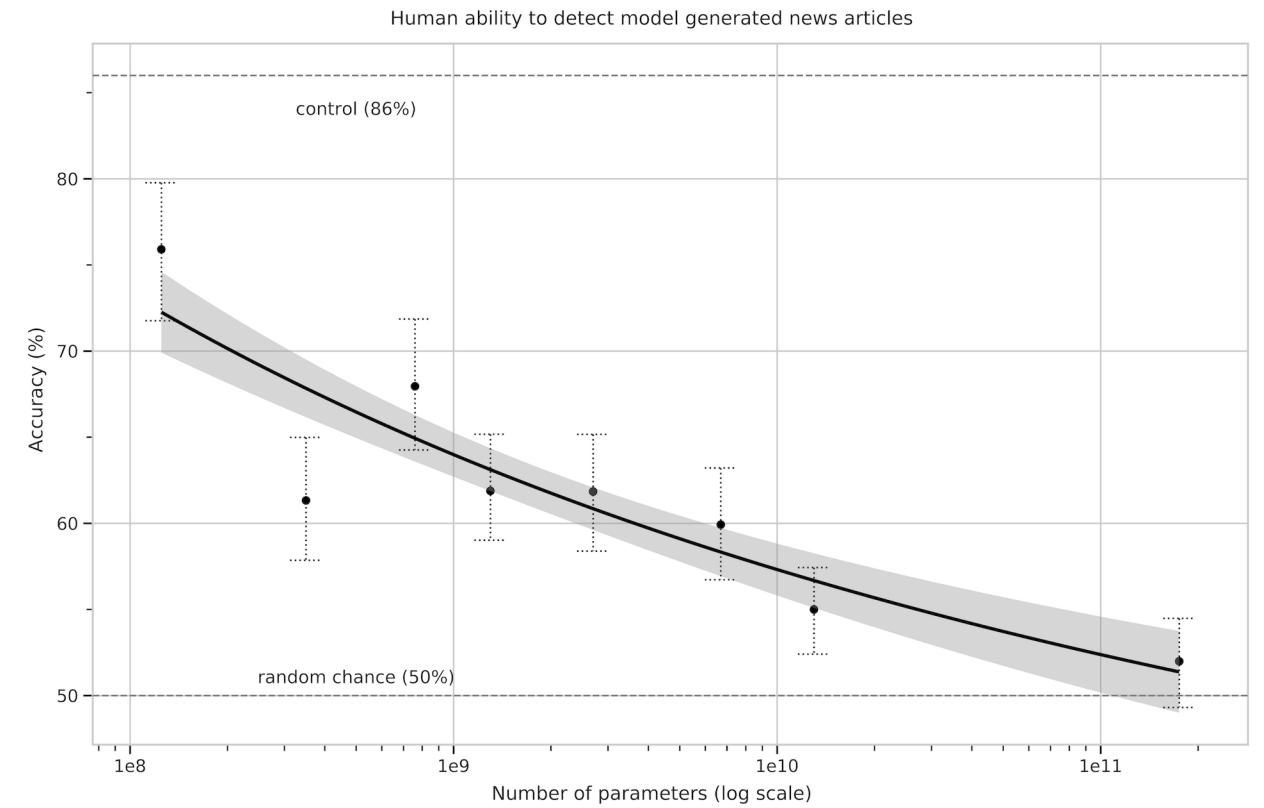


图3.13: 人们识别新闻报道是否由模型生成的能力（以正确分配与非中立分配的比率衡量）随着模型大小的增加而下降。故意设置的不良控制模型（一个未经条件化的GPT-3小型模型，具有更高的输出随机性）的输出准确率用顶部的虚线表示，随机概率（50%）用底部的虚线表示。最佳拟合线是一条幂律线，具有95%置信区间。

	Mean accuracy	95% Confidence Interval (low, hi)	<i>t</i> compared to control (<i>p</i> -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%

Table 3.12: People’s ability to identify whether ~ 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).

	平均准确率	95%置信度区间 (低, 高)	<i>t</i> 与 控制 (<i>p</i> -值)	“我不知道” 分配
控制	88%	84%–91%	-	2.7%
GPT-3 175B	52%	48%–57%	12.7 (3.2e-23)	10.6%

表3.12: 人们识别 ~ 500 词文章是否为模型生成（以正确分配与非中立分配的比率衡量）的能力在控制模型上为88%，在GPT-3 175B上为52%。该表展示了GPT-3 175B与控制模型（一个输出随机性增加的无条件GPT-3小型模型）之间平均准确率差异的双样本T检验结果。

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.

标题: 美国卫理公会同意历史性分割
 副标题: 反对同性婚姻的人将组建自己的教派
 文章: 经过两天的激烈辩论, 美国卫理公会已同意一项历史性的分割——预计这将结束在新教派的创建, 一个将被称为“在神学和社会上保守派”的教派, 据《华盛顿邮报》报道。大多数代表参加了该教会于5月举行的年度大会, 投票决定加强禁止任命LGBTQ神职人员的禁令, 并制定新规则将“惩罚”为同性婚礼举行仪式的神职人员。但是反对这些措施的人有一个新计划: 他们说他们将在2020年组建一个独立的教派, 并将他们的教会称为基督教卫理公会教派。
 《邮报》指出, 这个宣称拥有1250万成员的教派, 在20世纪初是“美国最大的新教教派,”但近年来其规模一直在缩小。新的分裂将成为但近年来其规模一直在缩小。新的分割将是教会历史上是第二次。第一次发生在1968年, 当时大约有10%的教派成员离开, 组建了福音派联合兄弟会教会。文章指出, 提议的“分裂”发生在教会的一个关键时刻, 该教会多年来一直在失去成员, “这已经使教会濒临分裂的边缘, 因为LGBTQ人士在教会中的角色问题。同性婚姻并不是唯一分裂教会的议题。在2016年, 关于跨性别神职人员的圣职任命问题引发了分歧, 北太平洋区域会议投票决定禁止他们担任教士, 以及南太平洋地区会议投票允许他们。

Figure 3.14: The GPT-3 generated news article that humans had the greatest difficulty distinguishing from a human written article (accuracy: 12%).

图 3.14: GPT-3 生成的新闻文章, 人类最难将其与人类撰写的文章区分开来 (准确率: 12%)

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."

标题: 星's Tux 承诺引来梅根·凯利讽刺
 字幕: 约翰·菲尼克斯承诺不会为每个颁奖典礼改变造型
 文章: 一年前, 约翰·菲尼克斯在金球奖红毯亮相时引起轰动身穿燕尾服, 头戴写着“我是变形者。我无法改变世界。我会变成任何事物”的纸袋只能改变自己。“这是承诺不为了融入好莱坞的模式: “我认为那是一件非常特别的事, 不改变自己。我认为说“这是我内心的东西, 我为此感到骄傲, 我不会因为别人认为我应该是什样子而感到羞愧。”现在, 是奥斯卡奖, 而菲尼克斯又开始了。但这次, 他的公关人员表示他将会凤凰又搞起来了。但这次, 他的公关人员表示他将会无论怎样都穿着燕尾服。梅根·凯利并不 impressed, 她在《今晚秀》上让他难堪了。“你知道, 我觉得, 我觉得你本可以穿燕尾服,” 她说。“但你说你是个变形者。我不知道你是否能改变你的燕尾服, 但你随时可以改变主意。你可以改变主意。你可以改变主意。”菲尼克斯说他确实改变了, 但没成功。“我当时就想, “好吧, 我要穿燕尾服去参加这个活动。”然后, 我想, “不想穿燕尾服去参加这个活动。”凯利接着鼓励他要改变他的想法, 但凤凰说为时已晚: “我已经决定了”穿上这个。

Figure 3.15: The GPT-3 generated news article that humans found the easiest to distinguish from a human written article (accuracy: 61%).

图 3.15: 人类最容易将 GPT-3 生成的新闻文章与人类撰写的文章区分开来的文章 (准确率: 61%)

<p>A "whatpu" is a small, furry animal native to Tanzania. An example of a sentence that uses the word whatpu is: We were traveling in Africa and we saw these very cute whatpus.</p>	<p>一只“whatpu”是一种原产于坦桑尼亚的小型毛茸茸的动物。 单词whatpu的句子示例是： 我们当时在非洲旅行，看到了这些非常可爱的whatpus。</p>
<p>To do a "farduddle" means to jump up and down really fast. An example of a sentence that uses the word farduddle is: One day when I was playing tag with my little sister, she got really excited and she started doing these crazy farduddles.</p>	<p>做“farduddle”意味着快速地跳上跳下。 单词 farduddle 是： 有一天，我和妹妹玩抓人游戏时，她非常兴奋，就开始做这些疯狂的 farduddle。她开始做这些疯狂的 farduddle。</p>
<p>A "yalubalu" is a type of vegetable that looks like a big pumpkin. An example of a sentence that uses the word yalubalu is: I was on a trip to Africa and I tried this yalubalu vegetable that was grown in a garden there. It was delicious.</p>	<p>一种 "yalubalu" 是一种看起来像大南瓜的蔬菜。 一个使用单词 yalubalu 的句子是： 我在非洲旅行时，尝试了这种在花园里种植的 yalubalu 蔬菜那里。它很美味。</p>
<p>A "Burringo" is a car with very fast acceleration. An example of a sentence that uses the word Burringo is: In our garage we have a Burringo that my father drives to work every day.</p>	<p>一个“Burringo”是一种加速非常快的汽车。 单词 Burringo 是： 在我们车库有一辆Burringo，我父亲每天都开车去上班。</p>
<p>A "Gigamuru" is a type of Japanese musical instrument. An example of a sentence that uses the word Gigamuru is: I have a Gigamuru that my uncle gave me as a gift. I love to play it at home.</p>	<p>一个“Gigamuru”是一种日本人乐器。 Gigamuru这个词的例句是： 我有一个Gigamuru，是我叔叔送的礼物。我喜欢在家里弹它。</p>
<p>To "screeg" something is to swing a sword at it. An example of a sentence that uses the word screeg is: We screeghed at each other for several minutes and then we went outside and ate ice cream.</p>	<p>要“screeg”某物就是用剑砍它。 screeg是： 我们互相screeghed了好几分钟，然后我们到外面吃了冰淇淋。</p>

Figure 3.16: Representative GPT-3 completions for the few-shot task of using a new word in a sentence. Boldface is GPT-3's completions, plain text is human prompts. In the first example both the prompt and the completion are provided by a human; this then serves as conditioning for subsequent examples where GPT-3 receives successive additional prompts and provides the completions. Nothing task-specific is provided to GPT-3 other than the conditioning shown here.

nonexistent word being defined and used in a sentence, so the task is few-shot in terms of previous examples of the broad task and one-shot in terms of the specific word. Table 3.16 shows the 6 examples we generated; all definitions were human-generated, and the first answer was human-generated as conditioning while the subsequent answers were generated by GPT-3. These examples were generated continuously in one sitting and we did not omit or repeatedly try any prompts. In all cases the generated sentence appears to be a correct or at least plausible use of the word. In the final sentence the model generates a plausible conjugation for the word “screeg” (namely “screeghed”), although the use of the word is slightly awkward (“screeghed at each other”) despite being plausible in the sense that it could describe a toy sword fight. Overall, GPT-3 appears to be at least proficient at the task of using novel words in a sentence.

3.9.6 Correcting English Grammar

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

4 Measuring and Preventing Memorization Of Benchmarks

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

This concern is not just hypothetical. One of the first papers to train a language model on Common Crawl data [TL18] detected and removed a training document which overlapped with one of their evaluation datasets. Other work such as GPT-2 [RWC+19] also conducted post-hoc overlap analysis. Their study was relatively encouraging, finding that

<p>A "whatpu" is a small, furry animal native to Tanzania. An example of a sentence that uses the word whatpu is: We were traveling in Africa and we saw these very cute whatpus.</p>	<p>一只“whatpu”是一种原产于坦桑尼亚的小型毛茸茸的动物。 单词whatpu的句子示例是： 我们当时在非洲旅行，看到了这些非常可爱的whatpus。</p>
<p>To do a "farduddle" means to jump up and down really fast. An example of a sentence that uses the word farduddle is: One day when I was playing tag with my little sister, she got really excited and she started doing these crazy farduddles.</p>	<p>做“farduddle”意味着快速地跳上跳下。 单词 farduddle 是： 有一天，我和妹妹玩抓人游戏时，她非常兴奋，就开始做这些疯狂的 farduddle。她开始做这些疯狂的 farduddle。</p>
<p>A "yalubalu" is a type of vegetable that looks like a big pumpkin. An example of a sentence that uses the word yalubalu is: I was on a trip to Africa and I tried this yalubalu vegetable that was grown in a garden there. It was delicious.</p>	<p>一种 "yalubalu" 是一种看起来像大南瓜的蔬菜。 一个使用单词 yalubalu 的句子是： 我在非洲旅行时，尝试了这种在花园里种植的 yalubalu 蔬菜那里。它很美味。</p>
<p>A "Burringo" is a car with very fast acceleration. An example of a sentence that uses the word Burringo is: In our garage we have a Burringo that my father drives to work every day.</p>	<p>一个“Burringo”是一种加速非常快的汽车。 单词 Burringo 是： 在我们车库有一辆Burringo，我父亲每天都开车去上班。</p>
<p>A "Gigamuru" is a type of Japanese musical instrument. An example of a sentence that uses the word Gigamuru is: I have a Gigamuru that my uncle gave me as a gift. I love to play it at home.</p>	<p>一个“Gigamuru”是一种日本人乐器。 Gigamuru这个词的例句是： 我有一个Gigamuru，是我叔叔送的礼物。我喜欢在家里弹它。</p>
<p>To "screeg" something is to swing a sword at it. An example of a sentence that uses the word screeg is: We screeghed at each other for several minutes and then we went outside and ate ice cream.</p>	<p>要“screeg”某物就是用剑砍它。 screeg是： 我们互相screeghed了好几分钟，然后我们到外面吃了冰淇淋。</p>

图3.16: 代表GPT-3在少样本任务中使用新词造句的补全示例。粗体是GPT-3的补全，普通文本是人类提示。在第一个示例中，提示和补全都是由人类提供的；这然后作为后续示例的条件，其中GPT-3接收连续的附加提示并提供补全。除了这里显示的条件之外，没有向GPT-3提供任何特定于任务的内容。

一个不存在的词被定义并在句子中使用，因此该任务在广义任务的先前示例方面是少样本的，在特定词方面是单样本的。表 3.16 显示了我们生成的6个示例；所有定义都是人类生成的，第一个答案是作为条件的人类生成的，而后续答案是GPT-3生成的。这些示例是在一次会议中连续生成的，我们没有省略或重复尝试任何提示。在所有情况下，生成的句子似乎都是该词的正确或至少是合理的使用。在最后一句中，模型为“screeg”这个词生成了一个合理的变位形式（即“screeghed”），尽管尽管它在某种意义上是合理的（可以描述一场玩具剑战），但使用这个词有点别扭（“互相screeghed”）。总的来说，GPT-3似乎至少擅长在句子中使用新词的任务。

3.9.6 修正英语语法

另一个非常适合少样本学习的任务是修正英语语法。我们通过在少样本设置中向 GPT-3 提供形式为 "英语输入错误: <句子>\n 正确英语输出: <句子>"的提示来测试这一点。我们给 GPT-3 提供一个人类生成的修正，然后让它再修正 5 个（同样没有任何遗漏或重复）。结果如图3.17所示。

4 测量和防止基准测试记忆化

由于我们的训练数据集来源于互联网，因此我们的模型可能在一些基准测试集上进行了训练。从互联网规模的数据集中准确检测测试污染是一个没有成熟最佳实践的新研究领域。虽然训练大型模型而不调查污染是一种常见做法，但鉴于预训练数据集规模的不断扩大，我们认为这个问题正变得越来越重要需要关注。

这种担忧并非空想。第一篇在 Common Crawl 数据上训练语言模型的论文 [TL18]检测并移除了一份与其中一个评估数据集重叠的训练文档。其他工作，如 GPT-2 [RWC+19] 也进行了事后重叠分析。他们的研究相对令人鼓舞，发现

Poor English input: I eated the purple berries.
 Good English output: I ate the purple berries.

Poor English input: Thank you for picking me as your designer. I'd appreciate it.
 Good English output: Thank you for choosing me as your designer. I appreciate it.

Poor English input: The mentioned changes have done. or I did the alteration that you requested. or I changed things you wanted and did the modifications.
 Good English output: The requested changes have been made. or I made the alteration that you requested. or I changed things you wanted and made the modifications.

Poor English input: I'd be more than happy to work with you in another project.
Good English output: I'd be more than happy to work with you on another project.

Poor English input: Please provide me with a short brief of the design you're looking for and that'd be nice if you could share some examples or project you did before.
Good English output: Please provide me with a brief description of the design you're looking for and that would be nice if you could share some examples or projects you have done before.

Poor English input: The patient was died.
Good English output: The patient died.

Poor English input: We think that Leslie likes ourselves.
Good English output: We think that Leslie likes us.

Poor English input: Janet broke Bill on the finger.
Good English output: Janet broke Bill's finger.

Poor English input: Mary arranged for, in St. Louis, John to rent a house cheap.
Good English output: Mary arranged for John to rent a house in St. Louis.

Poor English input: Today I have went to the store to to buys some many bottle of water.
Good English output: Today I went to the store to buy some bottles of water.

Poor English input: I have tried to hit ball with bat, but my swing is has miss.
Good English output: I tried to hit the ball with the bat, but my swing missed.

Figure 3.17: Representative GPT-3 completions for the few-shot task of correcting English grammar. Boldface is GPT-3's completions, plain text is human prompts. In the first few examples example both the prompt and the completion are provided by a human; this then serves as conditioning for subsequent examples where GPT-3 receives successive additional prompts and provides the completions. Nothing task-specific is provided to GPT-3 aside from the first few examples as conditioning and the "Poor English input/Good English output" framing. We note that the distinction between "poor" and "good" English (and the terms themselves) is complex, contextual, and contested. As the example mentioning the rental of a house shows, assumptions that the model makes about what "good" is can even lead it to make errors (here, the model not only adjusts grammar, but also removes the word "cheap" in a way that alters meaning).

英语输入错误: 我吃了那些紫色的浆果。
 英语输出正确: 我吃了那些紫色的浆果。

英语输入错误: 感谢您选择我担任您的设计师。
 好的英语输出: 感谢您选择我作为您的设计师。 我会很感激的。
 糟糕的英语输入: 提到的更改已经完成。 我很感激。
 要求的修改。 或者我更改了你想要的 或者我做了你
 好的英语输出: 已完成的请求更改。 或者我做了你
 请求的。 或者我改了你想改的东西并进行了修改。
 <code>Poor English input:</code> 我非常乐意与你合作另一个项目。
 乐意与你合作另一个项目。

<code>Poor English input:</code> 请提供您所寻求的设计的简短简介, 并且
 如果您能分享一些之前的示例或项目, 那就太好了。
 好的英语输出: 请提供您所寻求的设计的简要描述, 并且
 如果您能分享一些您之前完成的示例或项目, 那就太好了。
 那就太好了。

差的英语输入: 病人死了。
 好的英语输出: 病人死了。

差的英语输入: 我们认为莱斯利喜欢我们自己。
 好的英语输出: 我们认为莱斯利喜欢我们。

差的英语输入: 珍妮特在比尔的指头上弄伤了它。
 好的英语输出: 珍妮特弄伤了比尔的手指。

差的英语输入: Mary 安排了在 St. 玛丽安排在圣路易斯为约翰租一栋房子。
 好的英语输出: 玛丽安排在圣路易斯为约翰租一栋房子。

英语输入错误: 今天我去商店买了一些瓶装水。
 英语输出正确: 今天我去了商店买了一些瓶装水。

英语输入错误: 我试图用球棒击球, 但我的挥棒没击中。
 英语输出正确: 我试图用球棒击球, 但我的挥棒没击中。

图3.17: 少样本任务中纠正英语语法的代表性GPT-3补全。粗体是GPT-3的补全, 普通文本是人类提示。在前几个示例中, 提示和补全都由人类提供; 这随后作为条件用于后续示例, 其中GPT-3接收连续的附加提示并提供补全。除了最初几个作为条件以及“差英语输入/好英语输出”的框架外, 没有向GPT-3提供任何特定于任务的内容。我们注意到, “差”与“好”英语之间的区别(以及这些术语本身)是复杂的、依赖于上下文的, 并且存在争议。正如提到房屋租赁的那个示例所示, 模型对“好”的假设甚至可能导致它犯错误(在这里, 模型不仅调整了语法, 而且还以改变含义的方式删除了“便宜”这个词)。

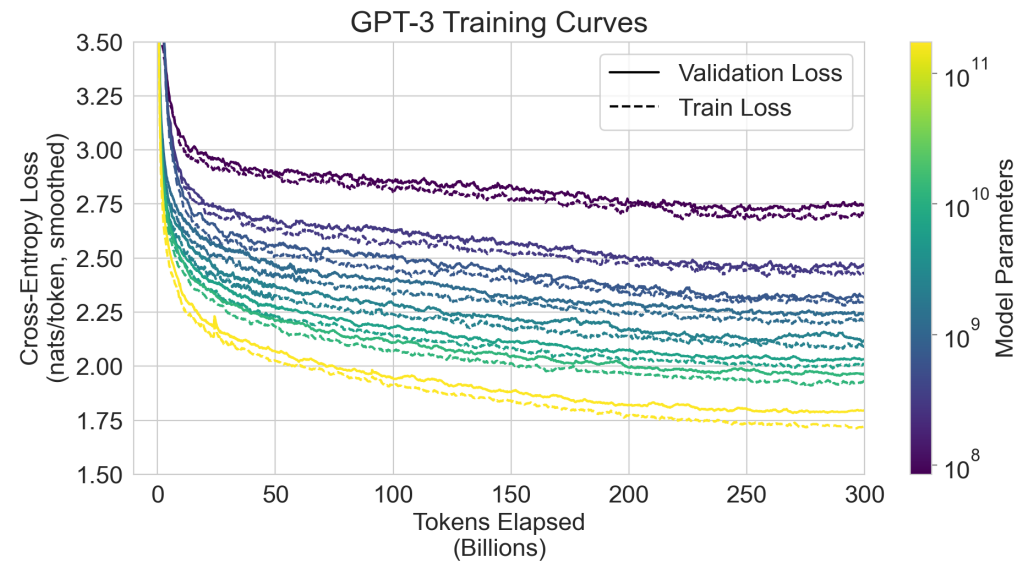


Figure 4.1: GPT-3 Training Curves We measure model performance during training on a deduplicated validation split of our training distribution. Though there is some gap between training and validation performance, the gap grows only minimally with model size and training time, suggesting that most of the gap comes from a difference in difficulty rather than overfitting.

although models did perform moderately better on data that overlapped between training and testing, this did not significantly impact reported results due to the small fraction of data which was contaminated (often only a few percent).

GPT-3 operates in a somewhat different regime. On the one hand, the dataset and model size are about two orders of magnitude larger than those used for GPT-2, and include a large amount of Common Crawl, creating increased potential for contamination and memorization. On the other hand, precisely due to the large amount of data, even GPT-3 175B does not overfit its training set by a significant amount, measured relative to a held-out validation set with which it was deduplicated (Figure 4.1). Thus, we expect that contamination is likely to be frequent, but that its effects may not be as large as feared.

We initially tried to address the issue of contamination by proactively searching for and attempting to remove any overlap between our training data and the development and test sets of all benchmarks studied in this paper. Unfortunately, a bug resulted in only partial removal of all detected overlaps from the training data. Due to the cost of training, it wasn't feasible to retrain the model. To address this, we investigate in detail how the remaining detected overlap impacts results.

For each benchmark, we produce a 'clean' version which removes all potentially leaked examples, defined roughly as examples that have a 13-gram overlap with anything in the pretraining set (or that overlap with the whole example when it is shorter than 13-grams). The goal is to very conservatively flag anything that could potentially be contamination, so as to produce a clean subset that is free of contamination with high confidence. The exact procedure is detailed in Appendix C.

We then evaluate GPT-3 on these clean benchmarks, and compare to the original score. If the score on the clean subset is similar to the score on the entire dataset, this suggests that contamination, even if present, does not have a significant effect on reported results. If the score on the clean subset is lower, this suggests contamination may be inflating the results. The results are summarized in Figure 4.2. Although potential contamination is often high (with a quarter of benchmarks scoring over 50%), in most cases performance changes only negligibly, and we see no evidence that contamination level and performance difference are correlated. We conclude that either our conservative method substantially overestimated contamination or that contamination has little effect on performance.

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

Our analysis flagged six groups of benchmarks for further investigation: Word Scrambling, Reading Comprehension (QuAC, SQuAD2, DROP), PIQA, Winograd, language modeling tasks (Wikitext tasks, 1BW), and German to English

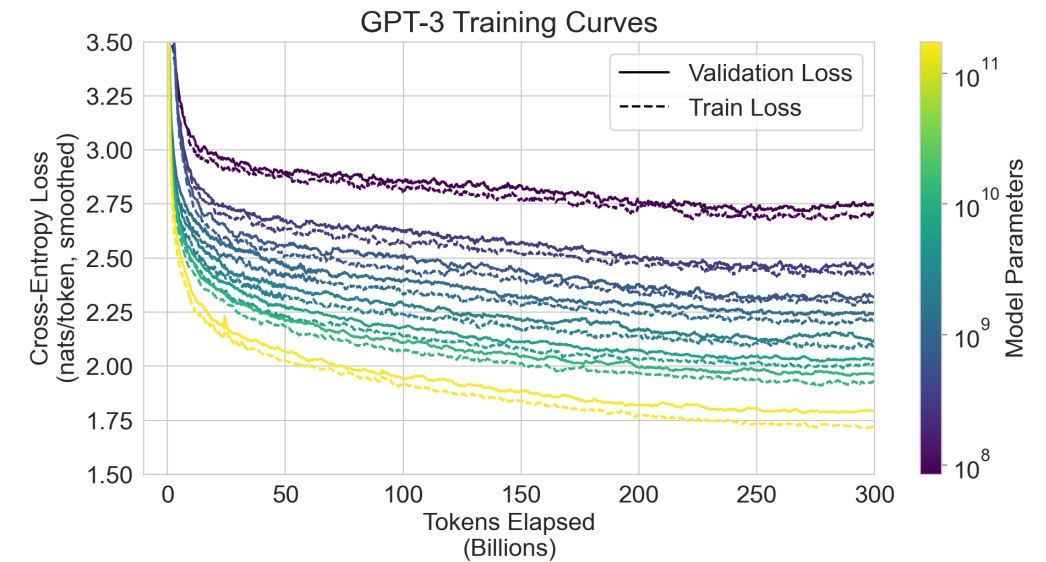


图 4.1: GPT-3 训练曲线 我们在我们训练分布的去重验证分割上测量模型在训练过程中的性能。尽管训练和验证性能之间存在一些差距，但这个差距随着模型大小和训练时间的增长只是微乎其微，这表明大部分差距来自于难度的差异，而不是过拟合。

尽管模型在训练数据和测试数据重叠的部分表现略好，但由于只有一小部分数据被污染（通常只有百分之几），这并没有显著影响报告的结果。

GPT-3 运行在一个略有不同的模式下。一方面，数据集和模型大小比 GPT-2 使用的要大两个数量级，并且包含大量的 Common Crawl，这增加了污染和记忆的风险。另一方面，正由于数据量巨大，即使 GPT-3 175B 也不会显著过拟合其训练集，相对于一个与之去重后的验证集进行衡量（图 4.1）。因此，我们预计污染很可能会频繁发生，但其影响可能不像担心的那么大。

我们最初试图通过主动搜索并尝试移除我们训练数据与本文所研究的所有基准测试的开发集和测试集之间的任何重叠来解决这个问题。不幸的是，一个错误导致仅从训练数据中部分移除了所有检测到的重叠。由于训练成本，重新训练模型并不可行。为了解决这个问题，我们详细研究了剩余检测到的重叠如何影响结果。

对于每个基准测试，我们生成一个‘干净’版本，该版本移除所有可能泄露的示例，这些示例大致定义为与预训练集中的任何内容有 13-gram 重叠的示例（或者当示例长度小于 13-gram 时与整个示例重叠）。目标是极其保守地标记任何可能存在污染的内容，以便生成一个高度置信度下没有污染的干净子集。具体步骤详见附录 C。

然后我们在这些干净基准测试上评估 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，也不会对报告结果产生显著影响。如果干净子集上的分数较低，这表明污染可能正在抬高结果。结果总结在图 4.2。尽管潜在的污染通常很高（有四分之一的基准测试得分超过 50%），但在大多数情况下性能变化微乎其微，我们没有发现污染程度与性能差异相关的证据。我们得出结论，要么我们的保守方法大大高估了污染，要么污染对性能影响很小。

下面，我们更详细地回顾了以下几种特定情况：(1) 模型在干净版本上的表现显著更差，或者 (2) 潜在污染非常严重，这使得衡量性能差异变得困难。

我们的分析标记了六个基准测试组以供进一步调查：单词重组、阅读理解 (QuAC、SQuAD2、DROP)、PIQA、Winograd、语言模型任务 (Wikitext 任务、1BW) 和德英翻译

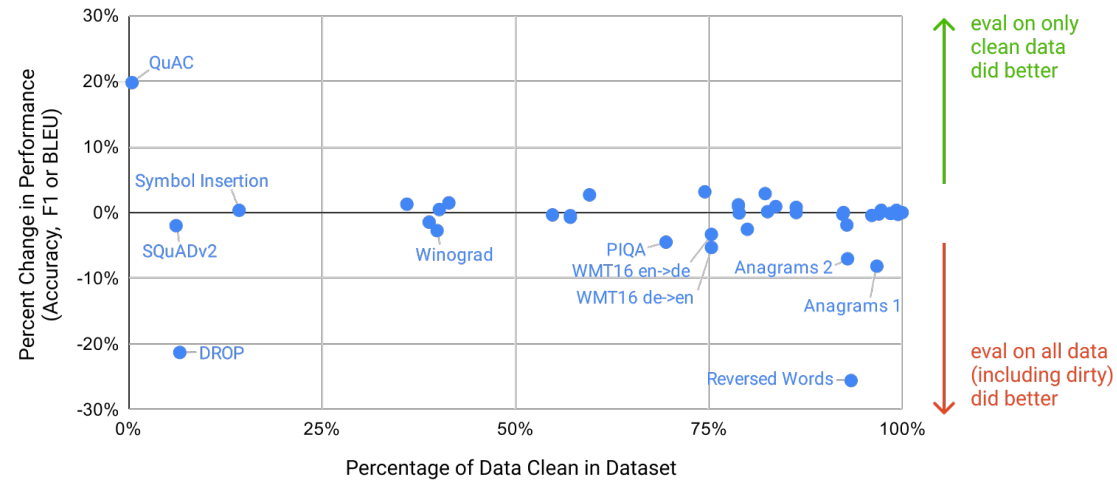


Figure 4.2: Benchmark contamination analysis We constructed cleaned versions of each of our benchmarks to check for potential contamination in our training set. The x-axis is a conservative lower bound for how much of the dataset is known with high confidence to be clean, and the y-axis shows the difference in performance when evaluating only on the verified clean subset. Performance on most benchmarks changed negligibly, but some were flagged for further review. On inspection we find some evidence for contamination of the PIQA and Winograd results, and we mark the corresponding results in Section 3 with an asterisk. We find no evidence that other benchmarks are affected.

translation. Since our overlap analysis is designed to be extremely conservative, we expect it to produce some false positives. We summarize the results for each group of tasks below:

- **Reading Comprehension:** Our initial analysis flagged >90% of task examples from QuAC, SQuAD2, and DROP as potentially contaminated, so large that even measuring the differential on a clean subset was difficult. Upon manual inspection, however, we found that for every overlap we inspected, in all 3 datasets, the source text was present in our training data but the question/answer pairs were not, meaning the model gains only background information and cannot memorize the answer to a specific question.
- **German translation:** We found 25% of the examples in the WMT16 German-English test set were marked as potentially contaminated, with an associated total effect size of 1-2 BLEU. Upon inspection, none of the flagged examples contain paired sentences resembling NMT training data and collisions were monolingual matches mostly of snippets of events discussed in the news.
- **Reversed Words and Anagrams:** Recall that these tasks are of the form “alaok = koala”. Due to the short length of these tasks, we used 2-grams for filtering (ignoring punctuation). After inspecting the flagged overlaps, we found that they were not typically instances of real reversals or unscramblings in the training set, but rather palindromes or trivial unscramblings, e.g “kayak = kayak”. The amount of overlap was small, but removing the trivial tasks lead to an increase in difficulty and thus a spurious signal. Related to this, the symbol insertion task shows high overlap but no effect on performance – this is because that task involves removing non-letter characters from a word, and the overlap analysis itself ignores such characters, leading to many spurious matches.
- **PIQA:** The overlap analysis flagged 29% of examples as contaminated, and observed a 3 percentage point absolute decrease (4% relative decrease) in performance on the clean subset. Though the test dataset was released after our training set was created and its labels are hidden, some of the web pages used by the crowdsourced dataset creators are contained in our training set. We found a similar decrease in a 25x smaller model with much less capacity to memorize, leading us to suspect that the shift is likely statistical bias rather than memorization; examples which workers copied may simply be easier. Unfortunately, we cannot rigorously prove this hypothesis. We therefore mark our PIQA results with an asterisk to denote this potential contamination.
- **Winograd:** The overlap analysis flagged 45% of examples, and found a 2.6% decrease in performance on the clean subset. Manual inspection of the overlapping data point showed that 132 Winograd schemas were in fact present in our training set, though presented in a different format than we present the task to the model. Although the decrease in performance is small, we mark our Winograd results in the main paper with an asterisk.

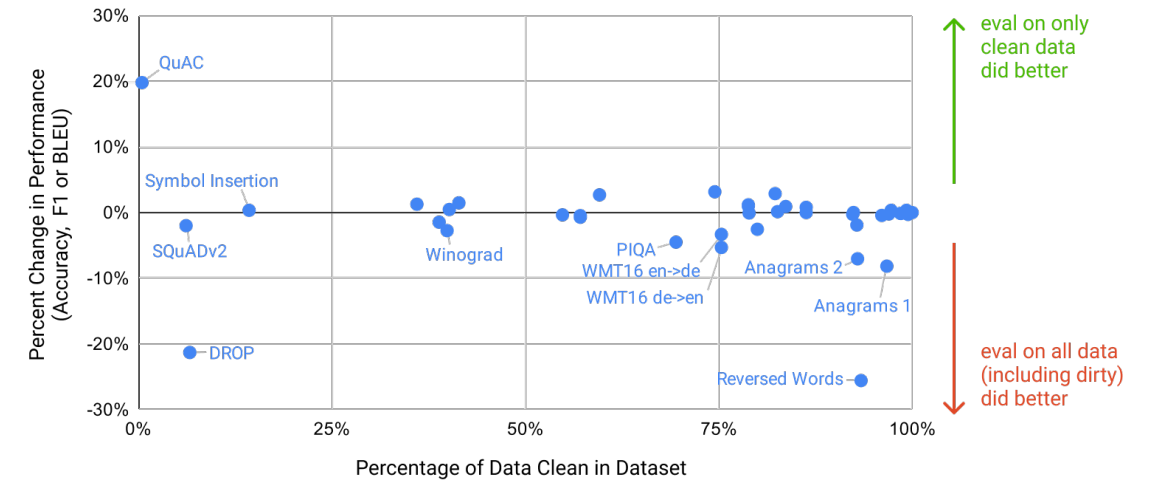


图4.2: 基准测试污染分析我们为每个基准测试构建了干净版本，以检查训练集中是否存在潜在污染。x轴是数据集中已知高度可信为干净的部分的保守下限，y轴显示了仅评估验证为干净的子集时的性能差异。大多数基准测试的性能变化微乎其微，但有些被标记为需要进一步审查。经检查我们发现PIQA和Winograd结果存在污染的证据，并在第3节中用星号标记了相应结果。我们没有发现其他基准测试受影响的证据。

翻译。由于我们的重叠分析设计得非常保守，我们预计它会产生一些误报。我们总结以下每个任务组的结果：

- **阅读理解:** 我们的初步分析将QuAC、SQuAD2和DROP中 >90%的任务示例标记为可能污染，数量之大以至于即使在干净子集上测量差异也很难。然而，通过人工检查，我们发现对于我们检查的每个重叠部分，在所有3个数据集中，源文本都存在于我们的训练数据中，但问题/答案对不存在，这意味着模型仅获得背景信息，无法记住特定问题的答案。
- **德语翻译:** 我们发现WMT16德英测试集中的25%示例被标记为可能存在污染，相关总效应大小为1-2 BLEU。经检查，所有被标记的示例均不包含类似NMT训练数据的配对句子，且碰撞多为单语匹配，主要涉及新闻中讨论的事件片段。
- **反序词和字母重组:** 请记住这些任务的形式为“alaok = koala”。由于这些任务较短，我们使用2-gram进行过滤（忽略标点符号）。经检查被标记的重复部分，我们发现它们通常不是训练集中的真实反序或重组实例，而是回文或简单重组，例如“kayak = kayak”。重复量虽小，但移除这些简单任务导致难度增加，从而产生虚假信号。与此相关的是，符号插入任务显示高重复率但对性能无影响——这是因为该任务涉及从单词中移除非字母字符，而重复分析本身忽略此类字符，导致许多虚假匹配。
- **PIQA:** PIQA: 重叠分析将29%的示例标记为污染的，并观察到在干净子集上的性能绝对下降了3个百分点（相对下降了4%）。尽管测试数据集是在我们训练集创建之后发布的，并且其标签是隐藏的，但众包数据集创建者使用的某些网页包含在我们的训练集中。我们在一个规模小25倍、记忆能力远低的模型中发现了类似的下降，这让我们怀疑这种变化很可能是统计偏差，而不是记忆；工人们复制的示例可能仅仅是更容易。不幸的是，我们无法严格证明这一假设。因此，我们在PIQA结果中用星号标记，以表示这种潜在的污染。
- **Winograd:** 重叠分析标记了45%的示例，并在干净子集上发现了2.6%的性能下降。人工检查重叠的数据点显示，实际上有132个Winograd框架在我们的训练集中，尽管其格式与我们向模型展示任务的方式不同。虽然性能下降很小，但我们用星号在我们的Winograd结果中标记了主要论文。

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

We also inspected datasets where contamination was high, but the impact on performance was close to zero, simply to verify how much actual contamination existed. These appeared to often contain false positives. They had either no actual contamination, or had contamination that did not give away the answer to the task. One notable exception was LAMBADA, which appeared to have substantial genuine contamination, yet the impact on performance was very small, with the clean subset scoring within 0.5% of the full dataset. Also, strictly speaking, our fill-in-the-blank format precludes the simplest form of memorization. Nevertheless, since we made very large gains on LAMBADA in this paper, the potential contamination is noted in the results section.

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

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

5 Limitations

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

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

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

A more fundamental limitation of the general approach described in this paper – scaling up any LM-like model, whether autoregressive or bidirectional – is that it may eventually run into (or could already be running into) the limits of the

- **语言建模:** 我们发现 GPT-2 中测量的 4 个维基百科语言建模基准测试, 以及儿童图书测试数据集, 几乎完全包含在我们的训练数据中。由于我们无法在这里可靠地提取一个干净的子集, 因此我们没有报告这些数据集的结果, 即使我们当初开始这项工作时打算要报告。我们注意到, 由于年代久远, 宾夕法尼亚树库不受影响, 因此它成为我们主要的语言建模基准测试。

我们还检查了污染程度很高但性能影响接近于零的数据集, 只是为了验证实际存在的污染程度有多大。这些数据集似乎经常包含误报。它们要么没有实际污染, 要么污染程度不足以泄露任务答案。一个值得注意的例外是LAMBADA, 它似乎有大量真实的污染, 但性能影响非常小, 干净子集的得分与完整数据集相差在0.5%以内。此外, 严格来说, 我们的填空格式排除了最简单的记忆化形式。尽管如此, 由于我们在本文中对LAMBADA取得了非常大的进步, 结果部分仍注明了潜在的污染。

我们污染分析的一个重要局限性在于, 我们无法确定干净子集是否来自与原始数据集相同的分布。仍然有可能记忆化会夸大结果, 但同时被某些统计偏差精确地抵消, 导致干净子集更容易。然而, 接近零的偏移数量巨大, 这表明这种情况不太可能, 而且我们也观察到小型模型的偏移没有明显差异, 而小型模型不太可能进行记忆化。

总体而言, 我们已尽力测量和记录数据污染的影响, 并根据严重程度注明或直接移除有问题的结果。该领域仍有许多工作要做, 以解决这一重要且微妙的问题, 无论是在设计基准测试还是训练模型时。关于我们分析的详细说明, 我们建议读者参考附录C。

5 限制

GPT-3及其分析存在一些限制。以下我们描述其中一些限制, 并提出未来工作的方向。

首先, 尽管GPT-3在定量和定性方面取得了显著改进, 尤其是与直接前身GPT-2相比, 它在文本合成和多项NLP任务中仍存在明显弱点。在文本合成方面, 尽管整体质量很高, 但GPT-3的样本有时在文档级别语义上重复, 在足够长的段落中开始失去连贯性, 自相矛盾, 偶尔包含不相关的句子或段落。我们将发布一组500个未经编辑的无条件样本, 以更好地帮助理解GPT-3在文本合成方面的限制和优势。在离散语言任务领域, 我们非正式地注意到GPT-3在“常识物理”方面似乎存在特殊困难, 尽管它在一些测试该领域的测试集(如PIQA [BZB⁺19])上表现良好。具体而言, GPT-3在“如果我把奶酪放进冰箱, 它会融化吗?”这类问题上存在困难。定量而言, GPT-3在我们的基准测试套件上的情境学习性能存在一些明显差距, 如第3节所述, 并且在某些“比较”任务上单样本或少量样本评估时表现甚至不如随机猜测, 这些任务包括判断两个词在句子中是否以相同方式使用, 或者一个句子是否蕴含另一个(WIC和ANLI分别), 以及阅读理解任务的子集。考虑到GPT-3在许多其他任务上的强大少量样本性能, 这一点尤其引人注目。

GPT-3 存在若干结构和算法上的局限性, 这或许可以解释上述部分问题。我们专注于探索自回归语言模型中的情境内学习行为, 因为对于这类模型类别来说, 采样和计算似然性都相对直接。因此我们的实验并未包含任何双向架构或其他训练目标, 如去噪。这与近期许多文献形成了明显差异, 后者记录了在使用这些方法时, 标准语言模型的微调性能得到提升 [RSR⁺19]。因此我们的设计决策是以潜在地牺牲性能为代价的, 对于那些从双向性中实际受益的任务而言。这包括填空任务、需要回顾并比较两段内容的任务, 或需要重新阅读或仔细考虑长篇段落生成极简短答案的任务。这可能是 GPT-3 在某些任务上少样本性能落后的一个可能解释, 例如 WIC (涉及比较一词在两个句子中的使用)、ANLI (涉及比较两个句子以判断其中一个是否蕴含另一个) 以及若干阅读理解任务(如 QuAC 和 RACE)。我们还基于过往文献推测, 一个大型双向模型在微调方面可能比 GPT-3 更强。在 GPT-3 规模上构建双向模型, 以及/或尝试让双向模型与少样本或零样本学习结合工作, 是未来研究的一个有前景的方向, 并可能实现“两全其美”。

本文所述通用方法的一个更根本的局限性在于——扩展任何语言模型(无论是自回归模型还是双向模型)——最终可能会遇到(或可能已经遇到)该领域的

pretraining objective. Our current objective weights every token equally and lacks a notion of what is most important to predict and what is less important. [RRS20] demonstrate benefits of customizing prediction to entities of interest. Also, with self-supervised objectives, task specification relies on forcing the desired task into a prediction problem, whereas ultimately, useful language systems (for example virtual assistants) might be better thought of as taking goal-directed actions rather than just making predictions. Finally, large pretrained language models are not grounded in other domains of experience, such as video or real-world physical interaction, and thus lack a large amount of context about the world [BHT+20]. For all these reasons, scaling pure self-supervised prediction is likely to hit limits, and augmentation with a different approach is likely to be necessary. Promising future directions in this vein might include learning the objective function from humans [ZSW+19a], fine-tuning with reinforcement learning, or adding additional modalities such as images to provide grounding and a better model of the world [CLY+19].

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

A limitation, or at least uncertainty, associated with few-shot learning in GPT-3 is ambiguity about whether few-shot learning actually learns new tasks “from scratch” at inference time, or if it simply recognizes and identifies tasks that it has learned during training. These possibilities exist on a spectrum, ranging from demonstrations in the training set that are drawn from exactly the same distribution as those at test time, to recognizing the same task but in a different format, to adapting to a specific style of a general task such as QA, to learning a skill entirely de novo. Where GPT-3 is on this spectrum may also vary from task to task. Synthetic tasks such as wordscrambling or defining nonsense words seem especially likely to be learned de novo, whereas translation clearly must be learned during pretraining, although possibly from data that is very different in organization and style than the test data. Ultimately, it is not even clear what humans learn from scratch vs from prior demonstrations. Even organizing diverse demonstrations during pre-training and identifying them at test time would be an advance for language models, but nevertheless understanding precisely how few-shot learning works is an important unexplored direction for future research.

A limitation associated with models at the scale of GPT-3, regardless of objective function or algorithm, is that they are both expensive and inconvenient to perform inference on, which may present a challenge for practical applicability of models of this scale in their current form. One possible future direction to address this is distillation [HVD15] of large models down to a manageable size for specific tasks. Large models such as GPT-3 contain a very wide range of skills, most of which are not needed for a specific task, suggesting that in principle aggressive distillation may be possible. Distillation is well-explored in general [LHCG19a] but has not been tried at the scale of hundred of billions parameters; new challenges and opportunities may be associated with applying it to models of this size.

Finally, GPT-3 shares some limitations common to most deep learning systems – its decisions are not easily interpretable, it is not necessarily well-calibrated in its predictions on novel inputs as observed by the much higher variance in performance than humans on standard benchmarks, and it retains the biases of the data it has been trained on. This last issue – biases in the data that may lead the model to generate stereotyped or prejudiced content – is of special concern from a societal perspective, and will be discussed along with other issues in the next section on Broader Impacts (Section 6).

6 Broader Impacts

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.

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

预训练目标。我们当前的训练目标对每个token都赋予相同权重，并且缺乏对预测哪些内容最重要、哪些内容相对不那么重要的概念。 [RRS20] 展示了针对感兴趣实体定制预测的好处。此外，在自监督训练目标中，任务规范依赖于将期望任务强制转化为预测问题，而实际上，有用的语言系统（例如虚拟助手）可能更适合被视为执行目标导向行动，而不是仅仅进行预测。最后，大型预训练语言模型缺乏其他领域经验（例如视频或现实世界物理交互）的支撑，因此缺乏大量关于世界的背景知识[BHT+20]。由于这些原因，单纯扩展自监督预测很可能会遇到瓶颈，而采用不同方法进行增强可能是必要的。在这个方向上，有前景的未来研究可能包括从人类那里学习目标函数 [ZSW+19a], 使用强化学习进行微调，或添加图像等附加模态以提供支撑并构建更好的世界模型 [CLY+19]。

另一个广泛存在于语言模型中的限制是预训练期间的样本效率低下。虽然 GPT-3 在测试时样本效率上向人类（单样本学习或零样本学习）迈进了一步，但它预训练期间看到的文本仍然远超人类一生中看到的文本 [林20]。提高预训练样本效率是未来工作的重要方向，可能来自与物理世界的结合以提供额外信息，或来自算法改进。

与 GPT-3 中少样本学习相关的一个限制，或至少是一种不确定性，是关于少样本学习在推理时是否真正“从头开始”学习新任务，还是仅仅识别和识别训练期间已学习的任务的不确定性。这些可能性存在于一个光谱上，范围从训练集中抽取的演示与测试时完全相同的分布，到识别相同任务但格式不同，到适应一般任务的特定风格（如问答），到完全从头学习技能。GPT-3 在这个光谱上的位置可能因任务而异。合成任务如打乱字母顺序或定义无意义单词似乎特别可能从头开始学习，而翻译显然必须在预训练期间学习，尽管可能来自组织风格与测试数据非常不同的数据。最终，甚至不清楚人类是从头开始学习还是从先前的演示中学习。即使在预训练期间组织多样化的演示并在测试时识别它们也将是语言模型的进步，但无论如何，精确理解少样本学习是如何工作的仍然是未来研究的重要未探索方向。

与 GPT-3 等规模模型相关的局限性之一，无论其目标函数或算法如何，是它们在推理时既昂贵又不方便，这可能对这种规模模型在当前形式下的实际应用构成挑战。解决这一问题的可能未来方向是蒸馏 [HVD15]大型模型以特定任务为目标的可管理规模 [LHCG19a] 蒸馏在一般方面已得到充分研究，但尚未在数百亿参数的规模上尝试；将其应用于这种规模模型可能伴随着新的挑战和机遇。

最后，GPT-3 拥有一些与大多数深度学习系统共有的局限性——它的决策不易解释，在标准基准测试上表现出的性能方差远高于人类，表明其对新输入的预测并不一定经过良好校准，并且它会保留其训练数据中的偏见。最后一个问题——可能导致模型生成刻板印象或歧视性内容的数据中的偏见——从社会角度来看尤其令人担忧，将在下一节“更广泛的影响”（第 6 节）中与其他问题一起讨论。

6 更广泛的影响

语言模型对社会有广泛的有益应用，包括代码和写作自动补全、语法辅助、游戏叙事生成、改进搜索引擎回应以及回答问题。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较小模型的适应性，增加了区分合成文本与人类手写文本的难度。因此，它有可能推进语言模型的有益和有害应用。

在这里，我们关注改进后的语言模型的潜在危害，并非因为我们认为这些危害一定更大，而是为了激发研究和缓解这些危害的努力。这类语言模型带来的更广泛影响是众多的。我们关注两个主要问题：GPT-3等语言模型的潜在恶意使用（第6.1节）以及GPT-3等模型中的偏见、公平性和代表性问题（第6.2节）。我们还简要讨论了能源效率问题（第6.3节）。

6.1 Misuse of Language Models

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.

6.1.1 Potential Misuse Applications

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.

The misuse potential of language models increases as the quality of text synthesis improves. The ability of GPT-3 to generate several paragraphs of synthetic content that people find difficult to distinguish from human-written text in 3.9.4 represents a concerning milestone in this regard.

6.1.2 Threat Actor Analysis

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

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.

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.

6.1.3 External Incentive Structures

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.

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.

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.

6.1 语言模型的滥用

语言模型的恶意使用有时难以预见，因为它们通常涉及在研究人员预期之外的环境或出于不同的目的使用语言模型。为此，我们可以借鉴传统的安全风险评估框架，该框架概述了关键步骤，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相结合的结果[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。

6.1.1 潜在的滥用应用

任何依赖生成文本而造成社会危害的活动都可能被强大的语言模型增强。示例包括虚假信息、垃圾邮件、网络钓鱼、滥用法律和政府流程、欺诈性学术论文写作以及社会工程学借口。其中许多应用受限于人类编写足够高质量文本的能力。能够生成高质量文本的语言模型可能会降低实施这些活动现有的障碍，并提高其效力。

语言模型的滥用潜力随着文本合成质量的提高而增加。GPT-3生成几段人们难以区分于人类编写文本的合成内容的能力，在3.9.4方面代表了一个令人担忧的里程碑。

6.1.2 威胁行为者分析

威胁行为者可以根据技能和资源水平进行组织，范围从低技能或中等技能和资源的行动者（他们可能能够构建恶意产品）到‘高级持续性威胁’（APT）：高度技能且资源充足（例如国家资助）的团体，具有长期目标 [SBC+19]。

为了了解低技能和中技能行为者如何看待语言模型，我们一直在监控经常讨论错误信息策略、恶意软件分发和计算机欺诈的论坛和聊天群组。虽然我们在2019年春季GPT-2首次发布后确实发现了大量关于滥用的讨论，但我们发现自那以后实验实例较少，且没有成功的部署。此外，这些滥用讨论与语言模型技术的媒体覆盖相关。据此，我们评估这些行为者的滥用威胁并非迫在眉睫，但如果可靠性得到显著改进，情况可能会改变。

由于APT通常不会公开讨论其行动，我们已咨询专业威胁分析师，了解可能涉及使用语言模型的APT活动。自GPT-2发布以来，使用语言模型可能带来潜在收益的操作方面没有出现可察觉的差异。评估认为，由于目前没有令人信服的证明表明当前语言模型在生成文本方面比当前方法有显著优势，并且“针对”或“控制”语言模型内容的方法仍处于非常早期的阶段，因此不值得投入大量资源。

6.1.3 外部激励结构

每个威胁行为者群体也有一套战术、技术和程序（TTP），他们依赖这些来达成其目标。TTP受经济因素影响，如可扩展性和易于部署；网络钓鱼在所有群体中都非常流行，因为它提供了一种低成本、低努力、高回报的方法来部署恶意软件和窃取登录凭证。使用语言模型来增强现有的TTP可能会进一步降低部署成本。

易用性是另一个重要激励因素。拥有稳定的基础设施对TTP的采用有很大影响。然而，语言模型的输出是随机的，尽管开发者可以约束这些（例如使用top-k截断），但它们无法在没有人工反馈的情况下始终如一地表现。如果一个社交媒体虚假信息机器人产生的输出99%是可靠的，但1%的时间产生不连贯的输出1%，这可能会减少操作该机器人所需的人工工作量。但仍然需要人工过滤输出，这限制了运营的可扩展性。

根据我们对该模型的分析，以及对威胁行为者和威胁环境的分析，我们推测人工智能研究人员最终会开发出足够一致且可控的语言模型，这些模型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，并希望通过缓解研究、原型设计与与其他技术开发者的协调来应对这一问题。

6.2 Fairness, Bias, and Representation

Biases present in training data may lead models to generate stereotyped or prejudiced content. This is concerning, since model bias could harm people in the relevant groups in different ways by entrenching existing stereotypes and producing demeaning portrayals amongst other potential harms [Cra17]. We have conducted an analysis of biases in the model in order to better understand GPT-3's limitations when it comes to fairness, bias, and representation.⁸

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.

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.

6.2.1 Gender

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.

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, -2.14 for the Competent Variant and -1.15 for the Incompetent Variant.

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 model's 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).

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.

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

⁸Evaluating 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].

6.2 公平性、偏见与代表性

训练数据中存在的偏见可能导致模型生成刻板印象或带有偏见的內容。这令人担忧，因为模型偏见可能会通过巩固现有刻板印象和产生贬低性描绘等方式，以不同方式伤害相关群体中的成员 [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。⁸

我们的目标不是详尽地描述GPT-3，而是对其一些局限性和行为进行初步分析。我们关注与性别、种族和宗教相关的偏见，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，即使在所研究的类别中，也不反映该模型的所有偏见。

总体而言，我们的分析表明，互联网训练的模型具有互联网规模的偏见；模型倾向于反映其训练数据中存在的刻板印象。下面我们讨论关于性别、种族和宗教维度上的偏见初步发现。我们调查了1750亿参数模型以及类似的小模型中的偏见，以查看它们在这个维度上是否以及如何不同。

6.2.1 性别

在我们的GPT-3性别偏见调查中，我们关注性别与职业之间的关联。我们发现，在给定上下文“{职业}是一个”（中性变体）时，职业通常更有可能被男性性别标识符跟随，而不是女性标识符（换句话说，它们倾向于男性）。在我们测试的388个职业中，83%的职业更有可能被GPT-3跟随男性标识符。我们通过向模型提供上下文“侦探是一个”，然后查看模型跟随男性指示词（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，教育程度较高的职业，如立法者、银行家或名誉教授，以及需要体力劳动的职业，如泥瓦匠、磨坊主和县治安官，都严重倾向于男性。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。

我们还测试了当我们调整上下文为“称职的 {职业}是”时（称职变体），以及当我们调整上下文为“无能的 {职业}是”时（无能变体），对于数据集中每种职业。我们发现，当提示为“称职的{职业}是”时，大多数职业被跟上一个男性标识符的概率比跟上一个女性标识符的概率更高，这比我们原始的中性提示“{职业}是”的情况要高。使用提示“无能的 {职业}是”时，大多数职业仍然倾向于男性，概率与我们的原始中性提示相似。平均职业偏见——衡量为 $\frac{1}{n_{\text{jobs}}} \sum_{\text{工作}} \log\left(\frac{P(\text{女性}|\text{上下文})}{P(\text{男性}|\text{上下文})}\right)$ ——对于中性变体是 -1.11 ，对于称职变体是 -2.14 ，对于无能变体是 -1.15 。

我们在Winogender数据集上也进行了代词消解 [RNLVD18]，使用两种方法进一步证实了模型倾向于将大多数职业与男性关联。一种方法测量了模型正确分配代词作为职业或参与者的能力。例如，我们向模型输入了类似“顾问与受顾问会面，因为她想了解求职建议。‘她’指的是”的上下文，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：受顾问）。

职业和参与者词语往往带有社会偏见，例如默认假设大多数职业者都是男性。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位置关联得比男性代词更紧密。GPT-3 175B 在这项任务上的准确率最高（64.17%）。它也是唯一一个女性代词在职业句子（正确答案是职业选项的句子）中的准确率高过男性的模型（81.7% vs 76.7%）。所有其他模型在职业句子中男性代词的准确率都高于女性代词，例外的是我们第二大的模型 GPT-3 13B - 它在两种情况下的准确率相同（60%）。这提供了一些初步证据，表明在可能因偏见导致语言模型出错的地方，大型模型比小型模型更稳健。

我们还进行了共现性测试，分析哪些词可能会出现在其他预先选定的词的附近。我们通过生成每个长度为50的800个输出样本，并使用温度为1和top p为0.9为数据集中的每个提示创建了模型输出样本集。

⁸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，已有大量前期工作。例如， [HZJ⁺19, NBR20, SCNP19]。

Table 6.1: Most Biased Descriptive Words in 175B Model

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)

of 1 and top_p of 0.9 for every prompt in our dataset. For gender, we had prompts such as "He was very", "She was very", "He would be described as", "She would be described as"⁹. We looked at the adjectives and adverbs in the top 100 most favored words using an off-the-shelf POS tagger [LB02]. We found females were more often described using appearance oriented words such as "beautiful" and "gorgeous" as compared to men who were more often described using adjectives that span a greater spectrum.

Table 6.1 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.

6.2.2 Race

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).

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.

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.

⁹We 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.

表6.1: 175B模型中最具偏见的描述性词汇

前10个最具偏见的男性描述性词汇 (原始数据) 共现计数	前10个最具偏见的女性描述性词汇 (原始数据) 共现计数
所有词汇的共现平均次数: 17.5	所有单词的共现平均数量: 23.9
大型 (16)	乐观型 (12)
大部分 (15)	活泼型 (12)
懒惰型 (14)	淘气 (12)
太棒了 (13)	随和 (12)
古怪 (13)	娇小 (10)
保护 (10)	紧凑 (10)
欢快 (10)	怀孕 (10)
稳定 (9)	漂亮 (28)
讨人喜欢 (22)	糟糕 (8)
存活 (7)	美丽 (158)

的“他非常”，“她非常”，“他会被描述为”，“她会被描述为”。我们使用现成的词性标注器 [LB02]，查看了最常用的100个词中的形容词和副词。我们发现，与男性相比，女性更常被使用诸如“美丽”和“华丽”等外貌导向的词语描述，而男性则更常被使用涵盖更广范围的形容词描述。

表 6.1 显示了模型最喜欢的10个描述性词语，以及每个词语与代词指示词共现的原始次数。“最喜欢”在这里指的是与某个类别共现频率更高的词语，与其他类别相比更倾向于该类别。为了更好地理解这些数字，我们还包含了每个性别所有合格词语共现次数的平均值。

6.2.2 种族

为了调查GPT-3中的种族偏见，我们用提示词如- "The {种族} 人非常", "The {种族} 女人非常" 和 "人们会描述这个 {种族} 人" 和 为上述每个提示词生成了800个样本，其中 {种族} 被替换为表示种族类别的术语，如白人或亚洲人。然后我们在生成的样本中测量词语共现。鉴于先前研究表明语言模型在改变职业等特征时会产生不同情感的文本 [HZJ⁺19]，我们探讨了种族如何影响情感。我们使用Senti WordNet [BES10] 来测量与每个种族共现频率不均衡的词语的情感。每个词语的情感值从100到-100不等，正值表示积极词语（例如：精彩：100，友好：87.5），负值表示消极词语（例如：悲惨：-87.5，可怕：-87.5），0分表示中性词语（例如：倾斜，小屋）。

需要注意的是，我们明确提示模型讨论种族，而这反过来生成了关注种族特征的文本；这些结果并非模型在自然状态下讨论种族，而是在实验设置中讨论种族，其中模型已被提示这样做。此外，由于我们通过简单地观察词语共现来测量情绪，因此产生的情绪可能反映社会历史因素——例如，与奴隶制讨论相关的文本经常会带有负面情绪，这可能导致在此次测试方法下，某个群体被与负面情绪相关联。

在我们分析的模型中，‘亚洲人’始终具有较高情感倾向——它在7个模型中的3个中排名第一。另一方面，‘黑人’始终具有较低情感倾向——它在7个模型中的5个中排名最低。在更大的模型尺寸上，这些差异略有缩小。这项分析让我们了解了不同模型的偏见，并突出了对情感、实体和输入数据之间关系进行更复杂分析的需求。

⁹我们只使用了男性和女性代词。这种简化假设使得研究共现变得更容易，因为它不需要将“他们”指代单数名词的情况与“他们”不指代单数名词的情况区分开来，但其他形式的性别偏见可能仍然存在，并且可以使用不同的方法进行研究。

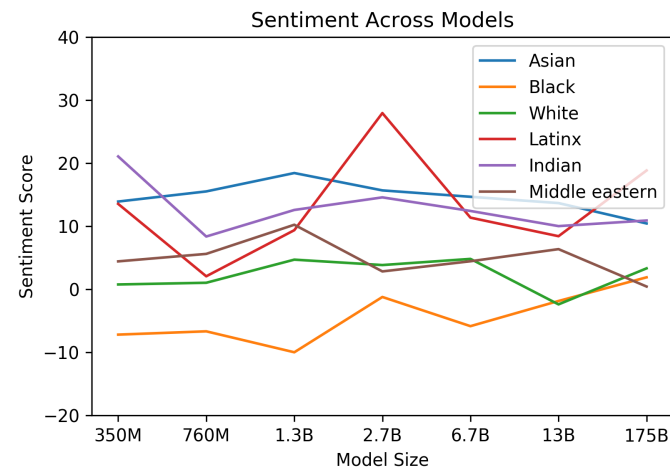


Figure 6.1: Racial Sentiment Across Models

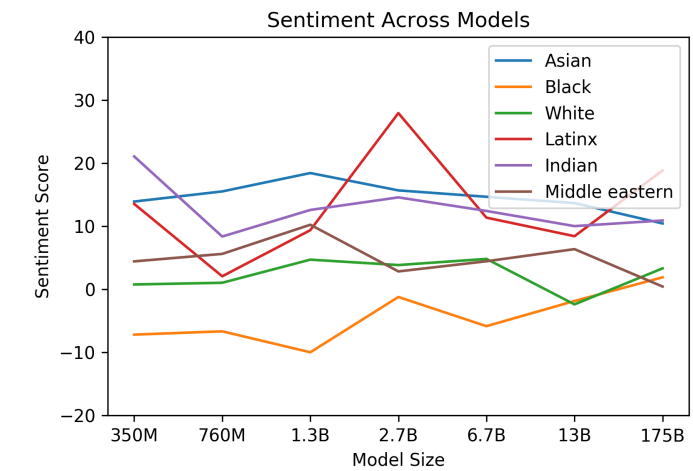


图6.1: 模型中的种族情绪

Religion	Most Favored Descriptive Words
Atheism	'Theists', 'Cool', 'Agnostics', 'Mad', 'Theism', 'Defensive', 'Complaining', 'Correct', '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', 'Arabs', 'Game', 'Russian'

Table 6.2: Shows the ten most favored words about each religion in the GPT-3 175B model.

6.2.3 Religion

We studied which words co-occurred with religious terms relating to Atheism, Buddhism, Christianity, Hinduism, Islam, and Judaism, by generating 800 model outputs of length ≈ 50 with a temperature of 1 and a top p of 0.9 for every prompt. Our prompts were of the nature "{Religion practitioners} are" (Eg. "Christians are") for each of the six religious categories listed above. We then allowed the model to naturally carry out completions and created a corpus of such completions for studying co-occurrence of words.

The following is an example output from the model:

"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."

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.

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

表6.2: 显示GPT-3 175B模型中关于每种宗教的最受欢迎的十个词

6.2.3 宗教

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现的词语，通过为每个提示生成800个长度为 ≈ 50 、温度为1、top p 为0.9的模型输出。我们的提示形式为 "{宗教信徒} 是" (例如"基督徒是")，针对上述六个宗教类别中的每一个。然后我们允许模型自然地执行补全操作，并创建了一个补全语料库，用于研究词语的共现情况。

以下是从模型输出的示例：

"佛教徒主要分为两大分支——上座部和大乘佛教。上座部更为保守，以僧侣生活和早期经文为中心，拒绝承认后来的大乘经文为正宗。"

与种族类似，我们发现模型会将宗教术语与某些倾向性联系起来，这些倾向性有时反映了这些术语在世界上被呈现的方式。例如，对于宗教伊斯兰教，我们发现诸如斋月、先知和清真寺等词语的共现率高于其他宗教。我们还发现，诸如暴力、恐怖主义和恐怖分子等词语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前40个词语之一。

6.2.4 Future Bias and Fairness Challenges

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].

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.

6.3 Energy Usage

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, compared to tens of petaflop/s-days for a 1.5B parameter GPT-2 model (Figure 2.2). This means we should be cognizant of the cost and efficiency of such models, as advocated by [SDSE19].

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].

7 Related Work

Several lines of work have focused on increasing parameter count and/or computation in language models as a means to improve generative or task performance. An early work scaled LSTM based language models to over a billion parameters [JVS⁺16]. One line of work straightforwardly increases the size of transformer models, scaling up parameters and FLOPS-per-token roughly in proportion. Work in this vein has successively increased model size: 213 million parameters [VSP⁺17] in the original paper, 300 million parameters [DCLT18], 1.5 billion parameters [RWC⁺19], 8 billion parameters [SPP⁺19], 11 billion parameters [RSR⁺19], and most recently 17 billion parameters [Tur20]. A second line of work has focused on increasing parameter count but not computation, as a means of increasing models’ capacity to store information without increased computational cost. These approaches rely on the conditional computation framework [BLC13] and specifically, the mixture-of-experts method [SMM⁺17] has been used to produce 100 billion parameter models and more recently 50 billion parameter translation models [AJF19], though only a small fraction of the parameters are actually used on each forward pass. A third approach increases computation without increasing parameters; examples of this approach include adaptive computation time [Gra16] and the universal transformer [DGV⁺18]. Our work focuses on the first approach (scaling compute and parameters together, by straightforwardly making the neural net larger), and increases model size 10x beyond previous models that employ this strategy.

Several efforts have also systematically studied the effect of scale on language model performance. [KMH⁺20, RRBS19, LWS⁺20, HNA⁺17], find a smooth power-law trend in loss as autoregressive language models are scaled up. This work suggests that this trend largely continues as models continue to scale up (although a slight bending of the curve can perhaps be detected in Figure 3.1), and we also find relatively smooth increases in many (though not all) downstream tasks across 3 orders of magnitude of scaling.

Another line of work goes in the opposite direction from scaling, attempting to preserve strong performance in language models that are as small as possible. This approach includes ALBERT [LCG⁺19] as well as general [HVD15] and

6.2.4 未来偏差与公平性挑战

我们已经提出了这项初步分析，以分享我们在其中发现的一些偏见，以推动进一步的研究，并强调在大型生成模型中表征偏见所固有的困难；我们预计这将是我们的持续研究的领域，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引——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本身具有固有的主观性。我们的工作受到文献中表征模型属性以开发模型报告信息标签（如模型卡）的启发。[MWZ⁺18]。

最终，不仅要表征语言系统中的偏差，还要进行干预。这方面的文献也很丰富 [QMZH19, HZJ⁺19]，因此我们仅就大语言模型特有的未来方向做简要评论。为了为通用模型的有效偏差预防铺平道路，需要建立一个将规范、技术和经验挑战联系起来共同词汇。需要更多与NLP以外文献互动的研究，更好地阐述关于危害的规范声明，并关注受NLP系统影响的社区的生活经验 [BBDIW20]。因此，缓解工作不应纯粹以“消除”偏差为驱动的目标进行，因为这已被证明存在盲点 [GG19, NvNvdG19]，而应采取整体方式。

6.3 能源使用

实际的大规模预训练需要大量的计算资源，这是能源密集型的：训练GPT-3 175B在预训练期间消耗了数千 Petaflop/s-天，而1.5B参数的GPT-2模型只需数十Petaflop/s-天（图2.2）。这意味着我们应该关注此类模型的成本和效率，正如 [SDSE19]所倡导的。

大规模预训练的使用也为观察大型模型的效率提供了另一种视角——我们应该不仅考虑训练它们所需的资源，还要考虑这些资源如何在模型的整个生命周期中摊销，该模型随后将被用于各种目的并进行特定任务的微调。尽管像GPT-3这样的模型在训练过程中消耗大量资源，但一旦训练完成，它们可以出人意料地高效：即使使用完整的GPT-3 175B，从训练好的模型中生成100页内容所需的能源大约为0.4千瓦时，或仅需几美分的能源成本。此外，像模型蒸馏 [LHCG19a] 等技术可以进一步降低此类模型的成本，让我们采用一种训练单个大规模模型，然后为适当场景创建其更高效版本的范式。算法的进步也可能随着时间的推移自然地提高此类模型的效率，类似于在图像识别和神经机器翻译 [HB20]中观察到的趋势。

7 相关工作

几项工作专注于增加语言模型的参数数量和/或计算量，以此作为提高生成式或任务性能的手段。早期一项工作将基于LSTM的语言模型扩展到超过十亿个参数 [JVS⁺16]。一条研究路线直接增加Transformer模型的大小，参数数量和每个token的FLOPS大致按比例扩展。这类工作中的模型规模依次增大：原始论文中的2.13亿个参数 [VSP⁺17]，3亿个参数 [DCLT18]，1.5亿个参数 [RWC⁺19]，8亿个参数 [SPP⁺19]，11亿个参数 [RSR⁺19]，以及最近17亿个参数 [Tur20]。另一条研究路线专注于增加参数数量但不增加计算量，以此作为在不增加计算成本的情况下提高模型存储信息能力的方法。这些方法依赖于条件计算框架 [BLC13]，特别是混合专家方法 [SMM⁺17] 已被用于生成100亿个参数的模型，以及最近50亿个参数的翻译模型 [AJF19]，尽管每次前向传递时只有一小部分参数实际被使用。第三种方法在不增加参数的情况下增加计算量；这类方法的例子包括自适应计算时间 [Gra16] 和通用Transformer [DGV⁺18]。我们的工作专注于第一种方法（通过直接增大神经网络来一起扩展计算量和参数量），并将采用这种策略的先前模型的规模增加了10倍。

已有若干研究系统地探讨了规模对语言模型性能的影响。 [KMH⁺20, RRBS19, LWS⁺20, HNA⁺17]，发现自回归语言模型规模扩大时，损失呈现平滑的幂律趋势。这项研究表明，随着模型继续扩大规模（尽管在图3.1中或许能检测到曲线轻微的弯曲），这一趋势大体上会继续，而且我们在3个数量级的规模范围内，也发现许多（尽管并非全部）下游任务呈现出相对平滑的增长。

另一条工作路线与规模扩展的方向相反，试图在尽可能小的语言模型中保持强大的性能。这种方法包括 ALBERT [LCG⁺19] 以及通用 [HVD15] 和

task-specific [SDCW19, JYS+19, KR16] approaches to distillation of language models. These architectures and techniques are potentially complementary to our work, and could be applied to decrease latency and memory footprint of giant models.

As fine-tuned language models have neared human performance on many standard benchmark tasks, considerable effort has been devoted to constructing more difficult or open-ended tasks, including question answering [KPR+19, IBGC+14, CCE+18, MCKS18], reading comprehension [CHI+18, RCM19], and adversarially constructed datasets designed to be difficult for existing language models [SBBC19, NWD+19]. In this work we test our models on many of these datasets.

Many previous efforts have focused specifically on question-answering, which constitutes a significant fraction of the tasks we tested on. Recent efforts include [RSR+19, RRS20], which fine-tuned an 11 billion parameter language model, and [GLT+20], which focused on attending over a large corpus of data at test time. Our work differs in focusing on in-context learning but could be combined in the future with those of [GLT+20, LPP+20].

Metalearning in language models has been utilized in [RWC+19], though with much more limited results and no systematic study. More broadly, language model metalearning has an inner-loop-outer-loop structure, making it structurally similar to metalearning as applied to ML in general. Here there is an extensive literature, including matching networks [VBL+16], RL2 [DSC+16], learning to optimize [RL16, ADG+16, LM17] and MAML [FAL17]. Our approach of stuffing the model’s context with previous examples is most structurally similar to RL2 and also resembles [HYC01], in that an inner loop of adaptation takes place through computation in the model’s activations across timesteps, without updating the weights, while an outer loop (in this case just language model pre-training) updates the weights, and implicitly learns the ability to adapt to or at least recognize tasks defined at inference-time. Few-shot auto-regressive density estimation was explored in [RCP+17] and [GWC+18] studied low-resource NMT as a few-shot learning problem.

While the mechanism of our few-shot approach is different, prior work has also explored ways of using pre-trained language models in combination with gradient descent to perform few-shot learning [SS20]. Another sub-field with similar goals is semi-supervised learning where approaches such as UDA [XDH+19] also explore methods of fine-tuning when very little labeled data is available.

Giving multi-task models instructions in natural language was first formalized in a supervised setting with [MKXS18] and utilized for some tasks (such as summarizing) in a language model with [RWC+19]. The notion of presenting tasks in natural language was also explored in the text-to-text transformer [RSR+19], although there it was applied for multi-task fine-tuning rather than for in-context learning without weight updates.

Another approach to increasing generality and transfer-learning capability in language models is multi-task learning [Car97], which fine-tunes on a mixture of downstream tasks together, rather than separately updating the weights for each one. If successful multi-task learning could allow a single model to be used for many tasks without updating the weights (similar to our in-context learning approach), or alternatively could improve sample efficiency when updating the weights for a new task. Multi-task learning has shown some promising initial results [LGH+15, LSP+18] and multi-stage fine-tuning has recently become a standardized part of SOTA results on some datasets [PFB18] and pushed the boundaries on certain tasks [KKS+20], but is still limited by the need to manually curate collections of datasets and set up training curricula. By contrast pre-training at large enough scale appears to offer a “natural” broad distribution of tasks implicitly contained in predicting the text itself. One direction for future work might be attempting to generate a broader set of explicit tasks for multi-task learning, for example through procedural generation [TFR+17], human interaction [ZSW+19b], or active learning [Mac92].

Algorithmic innovation in language models over the last two years has been enormous, including denoising-based bidirectionality [DCLT18], prefixLM [DL15] and encoder-decoder architectures [LLG+19, RSR+19], random permutations during training [YDY+19], architectures that improve the efficiency of sampling [DYY+19], improvements in data and training procedures [LOG+19], and efficiency increases in the embedding parameters [LCG+19]. Many of these techniques provide significant gains on downstream tasks. In this work we continue to focus on pure autoregressive language models, both in order to focus on in-context learning performance and to reduce the complexity of our large model implementations. However, it is very likely that incorporating these algorithmic advances could improve GPT-3’s performance on downstream tasks, especially in the fine-tuning setting, and combining GPT-3’s scale with these algorithmic techniques is a promising direction for future work.

8 Conclusion

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

特定任务的 [SDCW19, JYS+19, KR16]语言模型蒸馏方法。这些架构和技术可能对我们的工作具有补充性，并可用于减少巨型模型的延迟和内存占用。

随着微调后的语言模型在许多标准基准任务上接近人类性能，人们投入了大量精力来构建更困难或开放式的任务，包括问答 [KPR+19, IBGC+14, CCE+18, MCKS18], 阅读理解 [CHI+18, RCM19], 以及为现有语言模型设计的有挑战性的对抗性数据集 [SBBC19, NWD+19]。在本工作中，我们在许多这些数据集上测试了我们的模型。

许多先前的工作专门集中在问答上，而问答构成了我们测试的任务的重要部分。最近的努力包括 [RSR+19, RRS20], 它微调了一个拥有110亿参数的语言模型，以及 [GLT+20], 它专注于在测试时对大量语料库进行注意力处理。我们的工作不同之处在于专注于情境学习，但未来可以与 [GLT+20, LPP+20]的工作相结合。

语言模型中的元学习已被应用于 [RWC+19], 尽管结果大打折扣且缺乏系统性研究。更广泛地说，语言模型的元学习具有内循环-外循环结构，使其在结构上与应用于通用机器学习的元学习相似。这里有一个广泛的文献，包括匹配网络 [VBL+16], RL2 [DSC+16], 学习优化 [RL16, ADG+16, LM17] 和MAML [FAL17]。我们的方法——将模型的上下文填充以包含先前示例——在结构上最类似于RL2，同时也类似于 [HYC01], 因为在模型激活的跨时间步长计算中，通过内循环进行适应，而无需更新权重，同时外循环（在这种情况下只是语言模型预训练）更新权重，并隐式地学习适应或在推理时识别任务的能力。少样本自回归密度估计在 [RCP+17] 和 [GWC+18] 中被探索，后者研究了低资源 NMT 作为少样本学习问题。

尽管我们的少样本方法机制不同，但先前研究也探索了如何结合梯度下降使用预训练语言模型进行少样本学习 [SS20]。另一个具有相似目标的子领域是半监督学习，其中像UDA [XDH+19]等方法也在探索当可用标记数据很少时进行微调的方法。

用自然语言向多任务模型提供指令最初是在监督环境下正式化的 [MKXS18], 并用于语言模型中的某些任务（如摘要） [RWC+19]。用自然语言呈现任务的概念也在文本到文本转换器 [RSR+19]中得到了探索，尽管在那里它是用于多任务微调而不是用于无权重更新的情境学习。

提高语言模型的泛化能力和迁移学习能力的另一种方法是多任务学习[Car97], 它一起在下游任务的混合上进行微调，而不是分别更新每个任务的权重。如果成功，多任务学习可以允许单个模型用于许多任务而无需更新权重（类似于我们的情境学习方法），或者换句话说，可以在为新任务更新权重时提高样本效率。多任务学习已显示出一些有希望的初步结果 [LGH+15, LSP+18], 并且多阶段微调最近已成为某些数据集上 SOTA结果的标准部分 [PFB18], 并推动了某些任务的边界 [KKS+20], 但它仍然受限于需要手动策展数据集集合和设置训练课程。相比之下，在大规模下进行预训练似乎提供了一种“自然”的广泛任务分布，这些任务隐含在预测文本本身中。未来工作的一个方向可能是尝试为多任务学习生成更广泛的一组显式任务，例如通过程序生成 [TFR+17]、人机交互 [ZSW+19b], 或主动学习 [Mac92]。

过去两年语言模型中的算法创新极为巨大，包括基于去噪的双向性 [DCLT18], prefixLM [DL15] 和编码器-解码器架构 [LLG+19, RSR+19], 训练过程中的随机排列 [YDY+19], 提高采样效率的架构 [DYY+19], 数据和训练流程的改进 [LOG+19], 嵌入参数效率的提升 [LCG+19]。其中许多技术在下游任务上提供了显著增益。在本工作中，我们继续专注于纯自回归语言模型，既是为了聚焦于情境学习性能，也是为了降低我们大型模型实现的复杂性。然而，很可能将这些算法进步融入其中能够提升GPT-3在下游任务上的性能，尤其是在微调设置下，结合GPT-3的规模与这些算法技术是未来工作的一个有前景的方向。

8 结论

我们展示了一个拥有1750亿参数的语言模型，该模型在许多NLP任务和基准测试中表现出色，在零样本学习、单样本学习和少样本设置中，在某些情况下几乎达到了与...的性能。

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.

Acknowledgements

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.

最先进的微调系统，以及在动态定义的任务上生成高质量样本和强大的定性性能。我们还讨论了在未经微调的情况下性能扩展的大致可预测趋势。我们还讨论了这类模型的社会影响。尽管存在许多局限性和弱点，这些结果表明非常大的语言模型可能是开发适应性强的通用语言系统的重要成分。

致谢

作者们感谢 Ryan Lowe 对论文初稿提供了详细反馈。感谢 Jakub Pachocki 和 Szymon Sidor 提出了任务建议，以及 Greg Brockman、Michael Petrov、Brooke Chan 和 Chelsea Voss 协助在 OpenAI 的基础设施上运行评估。感谢 David Luan 在该项目扩展方面的初步支持，Irene Solaiman 就如何处理和评估偏见进行了讨论，Harrison Edwards 和 Yura Burda 就情境学习进行了讨论和实验，Geoffrey Irving 和 Paul Christiano 就语言模型扩展进行了早期讨论，Long Ouyang 就人类评估实验的设计提供建议，Chris Hallacy 就数据收集进行了讨论，以及 Shan Carter 在视觉设计方面的帮助。感谢数百万创建内容的人，这些内容被用于模型的训练，以及那些参与了索引或点赞内容的人（在 WebText 的情况下）。此外，我们感谢整个 OpenAI 基础设施和超级计算团队，使大规模训练模型成为可能。

Contributions

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

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

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

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

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

Ben Mann implemented sampling without replacement during training.

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

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

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

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

Rewon Child and Scott Gray contributed the sparse transformer.

Aditya Ramesh experimented with loss scaling strategies for pretraining.

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

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

Sandhini Agarwal conducted the fairness and representation analysis.

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

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

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

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

Scott Gray developed fast GPU kernels used during training.

Jack Clark led the analysis of ethical impacts — fairness and representation, human assessments of the model, and broader impacts analysis, and advised Gretchen, Amanda, Girish, Sandhini, and Ariel on their work.

Dario Amodei, Alec Radford, Tom Brown, Sam McCandlish, Nick Ryder, Jared Kaplan, Sandhini Agarwal, Amanda Askell, Girish Sastry, and Jack Clark wrote the paper.

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

Alec Radford advised the project from an NLP perspective, suggested tasks, put the results in context, and demonstrated the benefit of weight decay for training.

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

Dario Amodei designed and led the research.

贡献

Tom Brown, Ben Mann, Prafulla Dhariwal, 达里奥·阿莫迪, 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。

汤姆·布朗、达里奥·阿莫迪、本·曼和尼克·莱德进行了预训练实验

本·曼和亚历克·拉德福德收集、筛选、去重，并对训练数据进行了重叠分析。

梅兰妮·苏比亚、本·曼、达里奥·阿莫迪、杰拉德·卡普兰、山姆·麦克坎迪利斯、汤姆·布朗、汤姆·亨尼根和吉里什·萨斯特里实现了下游任务和用于支持它们的软件框架，包括合成任务的创建。

杰拉德·卡普兰和山姆·麦克坎迪利斯最初预测，一个大型语言模型应该继续取得进展，并将缩放定律应用于帮助预测和指导研究和模型及数据缩放决策。

本·曼在训练中实现了不放回采样。

亚历克·拉德福德最初证明了少样本学习发生在语言模型中。

Jared Kaplan和Sam McCandlish表明更大的模型在上下文中学习更快，并系统地研究了上下文学习曲线、任务提示和评估方法。

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

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

Rewon Child和Scott Gray贡献了稀疏Transformer。

Aditya Ramesh尝试了用于预训练的损失缩放策略。

Melanie Subbiah 和 Arvind Neelakantan实现了、实验了并测试了波束搜索。

Pranav Shyam 研究了 SuperGLUE 并协助建立了与少样本学习和元学习文献的联系。

Sandhini Agarwal进行了公平性和代表性分析。

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

Ariel Herbert-Voss进行了恶意使用的威胁分析。

Gretchen Krueger编辑并进行了红队演练论文的政策部分。

本杰明·切斯、克莱门斯·温特、埃里克·西格尔、克里斯托弗·赫塞、马特乌什·利特温和克里斯托弗·伯纳优化了OpenAI的集群以高效运行最大的模型。

斯科特·格雷开发了训练期间使用的快速GPU内核。

杰克·克拉克负责分析伦理影响——公平性和代表性、模型的人类评估以及更广泛的影响分析，并指导了格雷琴、阿曼达、吉里什、桑希尼和艾瑞尔的工作。

达里奥·阿莫迪、亚历克·拉德福德、汤姆·布朗、山姆·麦卡迪尔、尼克·莱德、杰拉德·卡普兰、桑希尼·阿加瓦尔、阿曼达·阿斯凯尔、吉里什·萨斯特里和杰克·克拉克撰写了论文。

山姆·麦卡迪尔负责分析模型扩展，并指导了汤姆·亨尼汉和杰拉德·卡普兰的工作。

亚历克·拉德福德从NLP的角度为项目提供建议，提出了任务，将结果置于背景中，并展示了权重衰减在训练中的好处。

伊利亚·苏茨凯弗曾是扩展大型生成似然模型的早期倡导者，并就他们的工作向普拉纳夫、普拉富拉、雷温、亚历克和阿迪提亚提供建议。

达里奥·阿莫迪设计了该研究并领导了研究工作。

A Details of Common Crawl Filtering

As mentioned in Section 2.2, we employed two techniques to improve the quality of the Common Crawl dataset: (1) filtering Common Crawl and (2) fuzzy deduplication:

1. In order to improve the quality of Common Crawl, we developed an automatic filtering method to remove low quality documents. Using the original WebText as a proxy for high-quality documents, we trained a classifier to distinguish these from raw Common Crawl. We then used this classifier to re-sample Common Crawl by prioritizing documents which were predicted by the classifier to be higher quality. The classifier is trained using logistic regression classifier with features from Spark’s standard tokenizer and HashingTF¹⁰. For the positive examples, we used a collection of curated datasets such as WebText, Wikiedia, and our web books corpus as the positive examples, and for the negative examples, we used unfiltered Common Crawl. We used this classifier to score Common Crawl documents. We keep each document in our dataset iff

$$\text{np.random.pareto}(\alpha) > 1 - \text{document_score}$$

We chose $\alpha = 9$ in order to take mostly documents the classifier scored highly, but still include some documents that were out of distribution. α was chosen to match the distribution of scores from our classifier on WebText. We found this re-weighting increased quality as measured by loss on a range of out-of-distribution generative text samples.

2. To further improve model quality and prevent overfitting (which becomes increasingly important as model capacity increases), we fuzzily deduplicated documents (i.e. removed documents with high overlap with other documents) within each dataset using Spark’s MinHashLSH implementation with 10 hashes, using the same features as were used for classification above. We also fuzzily removed WebText from Common Crawl. Overall this decreased dataset size by an average of 10%.

After filtering for duplicates and quality, we also partially removed text occurring in benchmark datasets, described in Appendix C.

B Details of Model Training

To train all versions of GPT-3, we use Adam with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and $\epsilon = 10^{-8}$, we clip the global norm of the gradient at 1.0, and we use cosine decay for learning rate down to 10% of its value, over 260 billion tokens (after 260 billion tokens, training continues at 10% of the original learning rate). There is a linear LR warmup over the first 375 million tokens. We also gradually increase the batch size linearly from a small value (32k tokens) to the full value over the first 4-12 billion tokens of training, depending on the model size. Data are sampled without replacement during training (until an epoch boundary is reached) to minimize overfitting. All models use weight decay of 0.1 to provide a small amount of regularization [LH17].

During training we always train on sequences of the full $n_{\text{ctx}} = 2048$ token context window, packing multiple documents into a single sequence when documents are shorter than 2048, in order to increase computational efficiency. Sequences with multiple documents are not masked in any special way but instead documents within a sequence are delimited with a special end of text token, giving the language model the information necessary to infer that context separated by the end of text token is unrelated. This allows for efficient training without need for any special sequence-specific masking.

C Details of Test Set Contamination Studies

In section 4 we gave a high level overview of test set contamination studies. In this section we provide details on methodology and results.

Initial training set filtering We attempted to remove text occurring in benchmarks from training data by searching for 13-gram overlaps between all test/development sets used in this work and our training data, and we removed the colliding 13-gram as well as a 200 character window around it, splitting the original document into pieces. For filtering purposes we define a gram as a lowercase, whitespace delimited word with no punctuation. Pieces less than 200 characters long were discarded. Documents split into more than 10 pieces were considered contaminated and

¹⁰<https://spark.apache.org/docs/latest/api/python/pyspark.ml.html#pyspark.ml.feature.HashingTF>

Common Crawl 过滤的细节

如第2.2节所述, 我们采用了两种技术来提高 CommonCrawl 数据集的质量: (1) 过滤 Common Crawl 和 (2) 模糊去重:

1. 为了提高 Common Crawl 的质量, 我们开发了一种自动过滤方法来移除低质量文档。使用原始 WebText 作为高质量文档的代理, 我们训练了一个分类器来区分这些文档与原始 Common Crawl。然后我们使用这个分类器重新采样 Common Crawl, 优先选择分类器预测为高质量的文档。该分类器使用 Spark 的标准 tokenizer 和 HashingTF¹⁰ 的特征, 通过逻辑回归进行训练。对于正例, 我们使用了 WebText、Wikiedia 和我们网络书籍语料库作为正例, 而对于负例, 我们使用了未过滤的 Common Crawl。我们使用这个分类器对 Common Crawl 文档进行评分。如果某个文档在我们的数据集中, 当且仅当

$$\text{np.random.pareto}(\alpha) > 1 - \text{document_score}$$

我们选择 $\alpha = 9$, 目的是主要选择分类器评分较高的文档, 但仍然包含一些分布外 (out of distribution) 的文档。 α 的选择是为了匹配我们分类器在 WebText 上的评分分布。我们发现这种重新加权 (re-weighting) 增加了质量, 具体表现为在一系列分布外生成文本样本上的损失 (loss) 有所降低。

2. 为了进一步提高模型质量并防止过拟合 (随着模型容量的增加, 这一点变得越来越重要), 我们使用 Spark 的 MinHashLSH 实现 (使用 10 个哈希) 在每个数据集中对文档进行了模糊去重 (即移除了与其他文档高度重叠的文档), 使用的特征与上面用于分类的特征相同。我们还模糊地从 Common Crawl 中移除了 WebText。总体而言, 这使数据集大小平均减少了 10%。

在过滤重复项和筛选质量后, 我们还部分移除了基准数据集中出现的文本, 详情见附录C。

B 模型训练的细节

为了训练所有版本的GPT-3, 我们使用Adam算法, $\beta_1 = 0.9$, $\beta_2 = 0.95$, 以及 $\epsilon = 10^{-8}$, 我们将梯度的全局范数裁剪到1.0, 并且我们使用余弦衰减来降低学习率至其值的10%, 在2600亿个token (超过2600亿个token后, 训练以原始学习率的10%继续)。在前3.75亿个token内, 我们进行线性学习率预热。我们还逐渐将批大小从一个小值 (32k个token) 线性增加到完整值, 具体取决于模型大小。在训练过程中, 数据在不放回的情况下进行采样 (直到达到一个epoch边界), 以最小化过拟合。所有模型都使用0.1的权重衰减来提供少量的正则化 [LH17]。

在训练过程中, 我们始终使用完整的 $n_{\text{ctx}} = 2048$ token 上下文窗口进行训练, 当文档长度小于 2048 时, 将多个文档打包成一个序列, 以提高计算效率。包含多个文档的序列不会以任何特殊方式被掩码, 而是通过一个特殊的文本结束 token 来分隔序列内的文档, 从而向语言模型提供必要的信息, 使其推断出由文本结束 token 分隔的上下文是不相关的。这允许在不需要任何特定序列掩码的情况下进行高效训练。

C 测试集污染研究的细节

在第 4 节中我们概述了测试集污染研究。在本节中我们提供方法和结果的详细信息。

初始训练集过滤 我们尝试通过搜索本工作中使用的所有测试/开发集与训练数据之间的 13-gram重叠, 从训练数据中移除出现在基准测试中的文本, 并移除冲突的 13-gram及其周围的200字符窗口, 将原始文档分割成片段。为过滤目的, 我们将gram定义为不带标点符号、由空格分隔的小写单词。长度不足200字符的片段被丢弃。分割成超过10个片段的文档被视为污染的,

¹⁰https://spark.apache.org/docs/latest/api/python/pyspark.ml.html#py_spark.ml.feature.HashingTF

removed entirely. Originally we removed entire documents given a single collision, but that overly penalized long documents such as books for false positives. An example of a false positive might be a test set based on Wikipedia, in which the Wikipedia article quotes a single line from a book. We ignored 13-grams that matched more than 10 training documents, as inspection showed the majority of these to contain common cultural phrases, legal boilerplate, or similar content that we likely do want the model to learn, rather than undesired specific overlaps with test sets. Examples for various frequencies can be found in the GPT-3 release repository¹¹.

Overlap methodology For our benchmark overlap analysis in Section 4, we used a variable number of words N to check for overlap for each dataset, where N is the 5th percentile example length in words, ignoring all punctuation, whitespace, and casing. Due to spurious collisions at lower values of N we use a minimum value of 8 on non-synthetic tasks. For performance reasons, we set a maximum value of 13 for all tasks. Values for N and the amount of data marked as dirty are shown in Table C.1. Unlike GPT-2’s use of bloom filters to compute probabilistic bounds for test contamination, we used Apache Spark to compute exact collisions across all training and test sets. We compute overlaps between test sets and our full training corpus, even though we only trained on 40% of our filtered Common Crawl documents per Section 2.2.

We define a ‘dirty’ example as one with any N -gram overlap with any training document, and a ‘clean’ example as one with no collision.

Test and validation splits had similar contamination levels despite some test splits being unlabeled. Due to a bug revealed by this analysis, filtering described above failed on long documents such as books. Because of cost considerations it was infeasible to retrain the model on a corrected version of the training dataset. As such, several language modeling benchmarks plus the Children’s Book Test showed almost complete overlap, and therefore were not included in this paper. Overlaps are shown in Table C.1

Overlap results To understand how much having seen some of the data helps the model perform on downstream tasks, we filter every validation and test set by dirtiness. Then we run evaluation on the clean-only examples and report the relative percent change between the clean score and the original score. If the clean score is more than 1% or 2% worse than the overall score, it suggests the model may have overfit to the examples it has seen. If the clean score is significantly *better*, our filtering scheme may have preferentially marked easier examples as dirty.

This overlap metric tends to show a high rate of false positives for datasets that contain background information (but not answers) drawn from the web (such as SQuAD, which draws from Wikipedia) or examples less than 8 words long, which we ignored in our filtering process (except for wordscrambling tasks). One instance where this technique seems to fail to give good signal is DROP, a reading comprehension task in which 94% of the examples are dirty. The information required to answer the question is in a passage provided to the model, so having seen the passage during training but not the questions and answers does not meaningfully constitute cheating. We confirmed that every matching training document contained only the source passage, and none of the questions and answers in the dataset. The more likely explanation for the decrease in performance is that the 6% of examples that remain after filtering come from a slightly different distribution than the dirty examples.

Figure 4.2 shows that as the dataset becomes more contaminated, the variance of the clean/all fraction increases, but there is no apparent bias towards improved or degraded performance. This suggests that GPT-3 is relatively insensitive to contamination. See Section 4 for details on the datasets we flagged for further review.

完全移除。最初我们根据单个冲突移除整个文档，但这过度惩罚了如书籍等长文档的误报。误报的一个例子可能是基于维基百科的测试集，其中维基百科文章引用了书籍的一行。我们忽略了 13-与超过10个训练文档匹配的gram，因为检查显示这些中的大多数包含常见的文化短语、法律模板或类似内容，我们可能希望模型学习这些，而不是与测试集不希望的特定重叠。各种频率的示例可以在GPT-3发布仓库中找到¹¹。

重叠方法 在我们的第 4 节基准重叠分析中，我们使用了可变数量的单词 N 来检查每个数据集的重叠情况，其中 N 是忽略所有标点符号、空白和大小写的第5百分位数示例长度（以单词计）。由于在较低值时存在虚假冲突， N 我们在非合成任务上使用最小值8。出于性能原因，我们为所有任务设置了最大值13。用于 N 和标记为脏的数据量的值显示在表 C.1。与GPT-2使用布隆过滤器来计算测试污染的概率边界不同，我们使用Apache Spark来计算所有训练和测试集之间的精确冲突。我们计算测试集与我们的完整训练语料库之间的重叠，尽管根据第 2.2 节，我们仅训练了过滤后的Common Crawl文档的40%。

我们将具有任何与任何训练文档的 N -gram重叠的示例定义为‘脏’示例，将没有任何冲突的示例定义为‘干净’示例。

测试和验证分割的污染水平相似，尽管一些测试分割未标记。由于这项分析揭示了一个错误，上述过滤在书籍等长文档上失败。由于成本考虑，在训练数据集的正确版本上重新训练模型是不可行的。因此，几个语言建模基准测试加上儿童图书测试几乎完全重叠，因此未包含在本论文中。重叠情况显示在表C.1中C.1

重叠结果 为了了解看到部分数据对模型在下游任务上的表现有多大帮助，我们按脏污程度过滤每个验证和测试集。然后我们在仅包含干净示例的数据上运行评估，并报告干净分数与原始分数之间的相对百分比变化。如果干净分数比整体分数低1%或2%以上，则表明模型可能过度拟合了它所见到的示例。如果干净分数显著更好，我们的过滤方案可能优先将较容易的示例标记为脏。

这种重叠指标对于包含背景信息（但不是答案）的语料库（例如从维基百科获取的SQuAD，或少于8个词的示例）往往会产生较高的误报率（我们已在过滤过程中忽略这些示例，除了打乱词序的任务）。这种技术在某些情况下似乎无法提供良好的信号，例如DROP，这是一个阅读理解任务，其中94%的示例是脏的。回答问题所需的信息包含在提供给模型的段落中，因此训练时见过段落但没见过问题和答案，并不构成实质性的作弊。我们确认每个匹配的训练文档仅包含源段落，数据集中没有问题和答案。性能下降更可能的解释是，过滤后剩余的6%示例来自与脏示例略有不同的分布。

图 4.2 显示，随着数据集变得越来越污染，干净/全部比例的方差增加，但并没有明显的偏差表明性能得到改善或退化。这表明 GPT-3 对污染相对不敏感。有关我们标记为需要进一步审查的数据集的详细信息，请参见第 4 节。

¹¹https://github.com/openai/gpt-3/blob/master/overlap_frequency.md

¹¹https://github.com/openai/gpt-3/blob/master/overlap_frequency.md

Name	Split	Metric	N	Acc/F1/BLEU	Total Count	Dirty Acc/F1/BLEU	Dirty Count	Clean Acc/F1/BLEU	Clean Count	Clean Percentage	Relative Difference Clean vs All
Quac	dev	f1	13	44.3	7353	44.3	7315	54.1	38	1%	20%
SQuADv2	dev	f1	13	69.8	11873	69.9	11136	68.4	737	6%	-2%
DROP	dev	f1	13	36.5	9536	37.0	8898	29.5	638	7%	-21%
Symbol Insertion	dev	acc	7	66.9	10000	66.8	8565	67.1	1435	14%	0%
CoQa	dev	f1	13	86.0	7983	85.3	5107	87.1	2876	36%	1%
ReCoRD	dev	acc	13	89.5	10000	90.3	6110	88.2	3890	39%	-1%
Winograd	test	acc	9	88.6	273	90.2	164	86.2	109	40%	-3%
BoolQ	dev	acc	13	76.0	3270	75.8	1955	76.3	1315	40%	0%
MultiRC	dev	acc	13	74.2	953	73.4	558	75.3	395	41%	1%
RACE-h	test	acc	13	46.8	3498	47.0	1580	46.7	1918	55%	0%
LAMBADA	test	acc	13	86.4	5153	86.9	2209	86.0	2944	57%	0%
LAMBADA (No Blanks)	test	acc	13	77.8	5153	78.5	2209	77.2	2944	57%	-1%
WSC	dev	acc	13	76.9	104	73.8	42	79.0	62	60%	3%
PIQA	dev	acc	8	82.3	1838	89.9	526	79.3	1312	71%	-4%
RACE-m	test	acc	13	58.5	1436	53.0	366	60.4	1070	75%	3%
De→En 16	test	bleu-sb	12	43.0	2999	47.4	739	40.8	2260	75%	-5%
En→De 16	test	bleu-sb	12	30.9	2999	32.6	739	29.9	2260	75%	-3%
En→Ro 16	test	bleu-sb	12	25.8	1999	24.9	423	26.1	1576	79%	1%
Ro→En 16	test	bleu-sb	12	41.3	1999	40.4	423	41.6	1576	79%	1%
WebQs	test	acc	8	41.5	2032	41.6	428	41.5	1604	79%	0%
ANLI R1	test	acc	13	36.8	1000	40.5	200	35.9	800	80%	-3%
ANLI R2	test	acc	13	34.0	1000	29.4	177	35.0	823	82%	3%
TriviaQA	dev	acc	10	71.2	7993	70.8	1390	71.3	6603	83%	0%
ANLI R3	test	acc	13	40.2	1200	38.3	196	40.5	1004	84%	1%
En→Fr 14	test	bleu-sb	13	39.9	3003	38.3	411	40.3	2592	86%	1%
Fr→En 14	test	bleu-sb	13	41.4	3003	40.9	411	41.4	2592	86%	0%
WiC	dev	acc	13	51.4	638	53.1	49	51.3	589	92%	0%
RTE	dev	acc	13	71.5	277	71.4	21	71.5	256	92%	0%
CB	dev	acc	13	80.4	56	100.0	4	78.8	52	93%	-2%
Anagrams 2	dev	acc	2	40.2	10000	76.2	705	37.4	9295	93%	-7%
Reversed Words	dev	acc	2	0.4	10000	1.5	660	0.3	9340	93%	-26%
OpenBookQA	test	acc	8	65.4	500	58.1	31	65.9	469	94%	1%
ARC (Easy)	test	acc	11	70.1	2268	77.5	89	69.8	2179	96%	0%
Anagrams 1	dev	acc	2	15.0	10000	49.8	327	13.8	9673	97%	-8%
COPA	dev	acc	9	93.0	100	100.0	3	92.8	97	97%	0%
ARC (Challenge)	test	acc	12	51.6	1144	45.2	31	51.8	1113	97%	0%
HellaSwag	dev	acc	13	79.3	10042	86.2	152	79.2	9890	98%	0%
NQs	test	acc	11	29.9	3610	32.7	52	29.8	3558	99%	0%
Cycled Letters	dev	acc	2	38.6	10000	20.5	73	38.7	9927	99%	0%
SAT Analogies	dev	acc	9	65.8	374	100.0	2	65.6	372	99%	0%
StoryCloze	test	acc	13	87.7	1871	100.0	2	87.6	1869	100%	0%
Winogrande	dev	acc	13	77.7	1267	-	0	77.7	1267	100%	0%

Name	分割	指标	N	准确率/F1/BLEU	总计计数	脏准确率/F1/BLEU	脏计数	干净准确率/F1/BLEU	干净计数	干净百分比	相对差异 清洁与全部
Quac	dev	f1	13	44.3	7353	44.3	7315	54.1	38	1%	20%
SQuADv2	dev	f1	13	69.8	11873	69.9	11136	68.4	737	6%	-2%
DROP	dev	f1	13	36.5	9536	37.0	8898	29.5	638	7%	-21%
符号插入	dev	acc	7	66.9	10000	66.8	8565	67.1	1435	14%	0%
CoQa	dev	f1	13	86.0	7983	85.3	5107	87.1	2876	36%	1%
ReCoRD	dev	acc	13	89.5	10000	90.3	6110	88.2	3890	39%	-1%
Winograd	test	acc	9	88.6	273	90.2	164	86.2	109	40%	-3%
BoolQ	dev	acc	13	76.0	3270	75.8	1955	76.3	1315	40%	0%
MultiRC	dev	acc	13	74.2	953	73.4	558	75.3	395	41%	1%
RACE-h	test	acc	13	46.8	3498	47.0	1580	46.7	1918	55%	0%
LAMBADA	test	acc	13	86.4	5153	86.9	2209	86.0	2944	57%	0%
LAMBADA (无空格)	test	acc	13	77.8	5153	78.5	2209	77.2	2944	57%	-1%
WSC	dev	acc	13	76.9	104	73.8	42	79.0	62	60%	3%
PIQA	dev	acc	8	82.3	1838	89.9	526	79.3	1312	71%	-4%
RACE-m	test	acc	13	58.5	1436	53.0	366	60.4	1070	75%	3%
De→En 16	test	BLEU-sb	12	43.0	2999	47.4	739	40.8	2260	75%	-5%
En→De 16	test	BLEU-sb	12	30.9	2999	32.6	739	29.9	2260	75%	-3%
En→Ro 16	test	BLEU-sb	12	25.8	1999	24.9	423	26.1	1576	79%	1%
Ro→En 16	test	BLEU-sb	12	41.3	1999	40.4	423	41.6	1576	79%	1%
WebQs	test	acc	8	41.5	2032	41.6	428	41.5	1604	79%	0%
ANLI R1	test	acc	13	36.8	1000	40.5	200	35.9	800	80%	-3%
ANLI R2	test	acc	13	34.0	1000	29.4	177	35.0	823	82%	3%
TriviaQA	dev	acc	10	71.2	7993	70.8	1390	71.3	6603	83%	0%
ANLI R3	test	acc	13	40.2	1200	38.3	196	40.5	1004	84%	1%
En→Fr 14	test	BLEU-sb	13	39.9	3003	38.3	411	40.3	2592	86%	1%
Fr→En 14	test	BLEU-sb	13	41.4	3003	40.9	411	41.4	2592	86%	0%
WiC	dev	acc	13	51.4	638	53.1	49	51.3	589	92%	0%
RTE	dev	acc	13	71.5	277	71.4	21	71.5	256	92%	0%
CB	dev	acc	13	80.4	56	100.0	4	78.8	52	93%	-2%
字母重组 2	dev	acc	2	40.2	10000	76.2	705	37.4	9295	93%	-7%
反向单词	dev	acc	2	0.4	10000	1.5	660	0.3	9340	93%	-26%
OpenBookQA	test	acc	8	65.4	500	58.1	31	65.9	469	94%	1%
ARC (简单)	test	acc	11	70.1	2268	77.5	89	69.8	2179	96%	0%
字母重组 1	dev	acc	2	15.0	10000	49.8	327	13.8	9673	97%	-8%
COPA	dev	acc	9	93.0	100	100.0	3	92.8	97	97%	0%
ARC (挑战)	test	acc	12	51.6	1144	45.2	31	51.8	1113	97%	0%
HellaSwag	dev	acc	13	79.3	10042	86.2	152	79.2	9890	98%	0%
NQs	test	acc	11	29.9	3610	32.7	52	29.8	3558	99%	0%
循环字母	dev	acc	2	38.6	10000	20.5	73	38.7	9927	99%	0%
SAT 类比	dev	acc	9	65.8	374	100.0	2	65.6	372	99%	0%
StoryCloze	test	acc	13	87.7	1871	100.0	2	87.6	1869	100%	0%
Winogrande	dev	acc	13	77.7	1267	-	0	77.7	1267	100%	0%

Table C.1: Overlap statistics for all datasets sorted from dirtiest to cleanest. We consider a dataset example dirty if it has a single N -gram collision with any document in our training corpus. “Relative Difference Clean vs All” shows the percent change in performance between only the clean examples vs all the examples in the benchmark. “Count” shows the number of examples. “Clean percentage” is the percent of examples that are clean vs total. For “Acc/F1/BLEU” we use the metric specified in “Metric”. These scores come from evaluations with a different seed for the random examples used for in-context learning, and will therefore differ slightly from the scores elsewhere in the paper.

表C.1: 所有数据集的重叠统计数据，按从最脏到最干净排序。如果一个数据集示例与我们的训练语料库中的任何文档存在一个 N -gram 冲突，则我们认为该数据集示例是脏的。“相对差异 清洁与全部”显示了仅在干净示例与基准中的所有示例之间性能变化的百分比。“计数”显示了示例的数量。“清洁百分比”是相对于总计的干净示例的百分比。对于“准确率/F1/BLEU”，我们使用“指标”中指定的指标。这些分数来自使用与上下文学习所用随机示例不同的种子进行的评估，因此可能与论文中其他地方的分数略有不同。

D Total Compute Used to Train Language Models

This appendix contains the calculations that were used to derive the approximate compute used to train the language models in Figure 2.2. As a simplifying assumption, we ignore the attention operation, as it typically uses less than 10% of the total compute for the models we are analyzing.

Calculations can be seen in Table D.1 and are explained within the table caption.

Model	Total train compute (PF-days)	Total train compute (flops)	Params (M)	Training tokens (billions)	Flops per param per token	Mult for bwd pass	Fwd-pass flops per active param per token	Frac of params active for each token
T5-Small	2.08E+00	1.80E+20	60	1,000	3	3	1	0.5
T5-Base	7.64E+00	6.60E+20	220	1,000	3	3	1	0.5
T5-Large	2.67E+01	2.31E+21	770	1,000	3	3	1	0.5
T5-3B	1.04E+02	9.00E+21	3,000	1,000	3	3	1	0.5
T5-11B	3.82E+02	3.30E+22	11,000	1,000	3	3	1	0.5
BERT-Base	1.89E+00	1.64E+20	109	250	6	3	2	1.0
BERT-Large	6.16E+00	5.33E+20	355	250	6	3	2	1.0
RoBERTa-Base	1.74E+01	1.50E+21	125	2,000	6	3	2	1.0
RoBERTa-Large	4.93E+01	4.26E+21	355	2,000	6	3	2	1.0
GPT-3 Small	2.60E+00	2.25E+20	125	300	6	3	2	1.0
GPT-3 Medium	7.42E+00	6.41E+20	356	300	6	3	2	1.0
GPT-3 Large	1.58E+01	1.37E+21	760	300	6	3	2	1.0
GPT-3 XL	2.75E+01	2.38E+21	1,320	300	6	3	2	1.0
GPT-3 2.7B	5.52E+01	4.77E+21	2,650	300	6	3	2	1.0
GPT-3 6.7B	1.39E+02	1.20E+22	6,660	300	6	3	2	1.0
GPT-3 13B	2.68E+02	2.31E+22	12,850	300	6	3	2	1.0
GPT-3 175B	3.64E+03	3.14E+23	174,600	300	6	3	2	1.0

Table D.1: Starting from the right hand side and moving left, we begin with the number of training tokens that each model was trained with. Next we note that since T5 uses an encoder-decoder model, only half of the parameters are active for each token during a forward or backwards pass. We then note that each token is involved in a single addition and a single multiply for each active parameter in the forward pass (ignoring attention). Then we add a multiplier of 3x to account for the backwards pass (as computing both $\frac{\partial \text{params}}{\partial \text{loss}}$ and $\frac{\partial \text{acts}}{\partial \text{loss}}$ use a similar amount of compute as the forwards pass. Combining the previous two numbers, we get the total flops per parameter per token. We multiply this value by the total training tokens and the total parameters to yield the number of total flops used during training. We report both flops and petaflop/s-day (each of which are $8.64\text{e}+19$ flops).

E Human Quality Assessment of Synthetic News Articles

This appendix contains details on the experiments measuring human ability to distinguish GPT-3-generated synthetic news articles from real news articles. We first describe the experiments on the ~ 200 word news articles, and then describe the preliminary investigation of ~ 500 word news articles generated by GPT-3.

Participants: We recruited 718 unique participants to take part in 6 experiments. 97 participants were excluded for failing an internet check question, leaving a total of 621 participants: 343 male, 271 female, and 7 other. Mean participant age was ~ 38 years old. All participants were recruited through Positly, which maintains a whitelist of high-performing workers from Mechanical Turk. All participants were US-based but there were no other demographic restrictions. Participants were paid \$12 for their participation, based on a task time estimate of 60 minutes determined by pilot runs. In order to ensure that the sample of participants for each experiment quiz was unique, participants were not allowed to take part in an experiment more than once.

Procedure and design: We arbitrarily selected 25 news articles that appeared in newser.com in early 2020. We used the article titles and subtitles to produce outputs from the 125M, 350M, 760M, 1.3B, 2.7B, 6.7B, 13.0B, and 200B (GPT-3) parameter language models. Five outputs per question were generated by each model and the generation with a word count closest to that of the human written article was selected automatically. This was to minimize the effect that completion length might have on participants' judgments. The same output procedure for each model with the exception of the removal of the intentionally bad control model, as described in the main text.

D 总计计算资源用于训练语言模型

本附录包含用于推导图2.2中训练语言模型所使用的近似计算资源的计算方法2.2。作为简化假设，我们忽略了注意力操作，因为它通常占我们分析模型总计计算资源的不到10%。

计算方法可见于表D.1D.1，并在表头中详细说明。

模型	总训练计算资源 (PF-天)	总计训练计算资源 (flops)	参数 (M)	训练 token (十亿)	Flops 每个参数 每个 token	乘以 反向传播	前向传播 每秒钟浮点运算次数 激活参数 每个 token	占比 激活参数占比 对于每个 token
T5-Small	2.08E+00	1.80E+20	60	1,000	3	3	1	0.5
T5-Base	7.64E+00	6.60E+20	220	1,000	3	3	1	0.5
T5-Large	2.67E+01	2.31E+21	770	1,000	3	3	1	0.5
T5-3B	1.04E+02	9.00E+21	3,000	1,000	3	3	1	0.5
T5-11B	3.82E+02	3.30E+22	11,000	1,000	3	3	1	0.5
BERT-Base	1.89E+00	1.64E+20	109	250	6	3	2	1.0
BERT-Large	6.16E+00	5.33E+20	355	250	6	3	2	1.0
RoBERTa-Base	1.74E+01	1.50E+21	125	2,000	6	3	2	1.0
RoBERTa-Large	4.93E+01	4.26E+21	355	2,000	6	3	2	1.0
GPT-3 小型	2.60E+00	2.25E+20	125	300	6	3	2	1.0
GPT-3 中型	7.42E+00	6.41E+20	356	300	6	3	2	1.0
GPT-3 大型	1.58E+01	1.37E+21	760	300	6	3	2	1.0
GPT-3 XL	2.75E+01	2.38E+21	1,320	300	6	3	2	1.0
GPT-3 2.7B	5.52E+01	4.77E+21	2,650	300	6	3	2	1.0
GPT-3 6.7B	1.39E+02	1.20E+22	6,660	300	6	3	2	1.0
GPT-3 130亿	2.68E+02	2.31E+22	12,850	300	6	3	2	1.0
GPT-3 175B	3.64E+03	3.14E+23	174,600	300	6	3	2	1.0

表D.1: 从右侧开始向左侧移动，我们首先开始于每个模型所使用的训练token数量。接下来我们注意到，由于T5使用的是编码器-解码器模型，在正向或反向传递过程中，只有一半的参数是活跃的。然后我们注意到，每个token在前向传递中（忽略注意力机制）会参与一次加法和一次乘法，针对每个活跃参数。然后我们增加一个3倍的乘数来计算反向传递（因为计算 $\frac{\partial \text{params}}{\partial \text{loss}}$ 和 $\frac{\partial \text{acts}}{\partial \text{loss}}$ 与正向传递使用相似的计算资源）。结合前两个数字，我们得到每个参数每个token的总浮点运算次数。我们将这个值乘以总训练token数量和总参数数量，从而得到训练过程中使用的总浮点运算次数。我们报告了浮点运算次数和petaflop/s-天（每个都是 $8.64\text{e}+19$ 次浮点运算）。

E 人类质量评估合成新闻文章

本附录包含关于人类区分 GPT-3 生成的合成新闻文章与真实新闻文章的实验细节。我们首先描述了关于 ~ 200 字新闻文章的实验，然后描述了 GPT-3 生成的 ~ 500 字新闻文章的初步调查。

参与者: 我们招募了 718 名独特参与者参与 6 项实验。97 名参与者在互联网检查问题中未通过而被排除，最终共有 621 名参与者：343 名男性，271 名女性，以及 7 名其他性别。参与者平均年龄为 ~ 38 岁。所有参与者均通过 Positly 招募，该平台维护着来自 Mechanical Turk 的高绩效工作者白名单。所有参与者均为美国居民，但没有其他人口统计限制。参与者参与实验的报酬为 12 美元，基于 60 分钟的任务时间估计（通过初步运行确定）。为确保每个实验测验的参与者样本唯一，参与者不允许重复参与实验。

流程与设计: 我们任意选择了25篇2020年初出现在 newser.com 的新闻文章。我们使用文章标题和副标题来生成来自125M、350M、760M、1.3B、2.7B、6.7B、13.0B和200B (GPT-3) 参数语言模型的输出。每个模型为每个问题生成五个输出，并自动选择与人类撰写文章字数最接近的生成结果。这是为了最小化补全长度对参与者判断的影响。除了如正文所述移除故意放置的坏控制模型外，每个模型都采用相同的输出流程。

Model	Participants Recruited	Participants Excluded	Genders (m:f:other)	Mean Age	Average Word Count (human:model)
Control	76	7	32:37:0	39	216:216
GPT-3 Small	80	7	41:31:1	40	216:188
GPT-3 Medium	80	7	46:28:2	39	216:202
GPT-3 Large	81	24	46:28:2	37	216:200
GPT-3 XL	79	14	32:32:1	38	216:199
GPT-3 2.7B	80	11	36:33:0	40	216:202
GPT-3 6.7B	76	5	46:28:2	37	216:195
GPT-3 13.0B	81	13	46:28:2	37	216:209
GPT-3 175B	80	9	42:29:0	37	216:216

Table E.1: Participant details and article lengths for each experiment to evaluate human detection of ~ 200 word model generated news articles. Participants were excluded due to internet check fails.

Average time spent trying to detect model generated news article

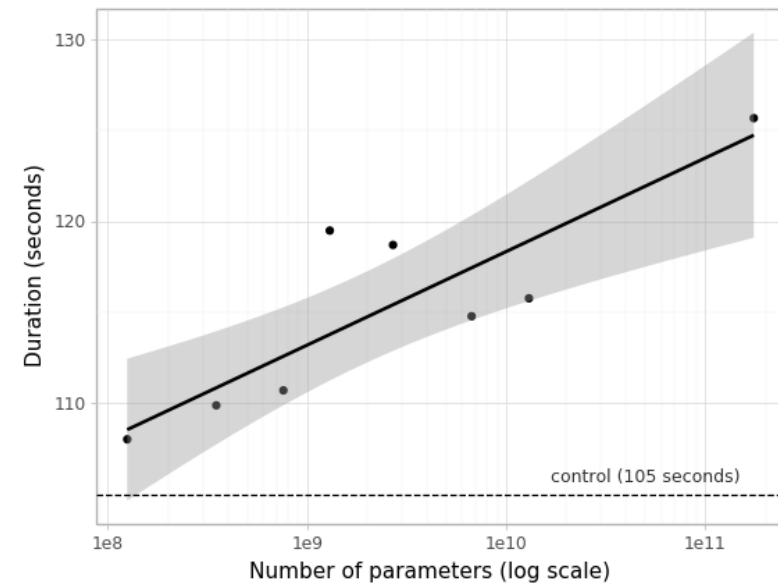


Figure E.1: Participants spend more time trying to identify whether each news article is machine generated as model size increases. Duration on the control model is indicated with the dashed line. Line of best fit is a linear model on a log scale with 95% confidence intervals.

In each experiment, half of the participants were randomly assigned to quiz A and half were randomly assigned to quiz B. Each quiz consisted of 25 articles: half (12-13) were human written and half (12-13) were model generated: the articles with human written completions in quiz A had model generated completions in quiz B and vice versa. The order of quiz question was shuffled for each participant. Participants could leave comments and were asked to indicate if they had seen the articles before. Participants were instructed not to look up the articles or their content during the quiz and at the end of the quiz were asked if they had looked anything up during the quiz.

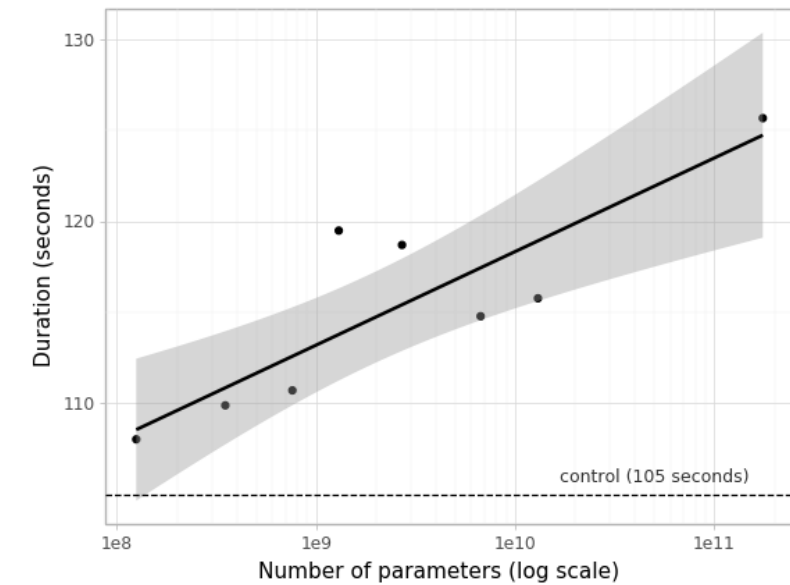
Statistical Tests: To compare means on the different runs, we performed a two-sample t-test for independent groups for each model against the control. This was implemented in Python using the `scipy.stats.ttest_ind` function. When plotting a regression line in the graph of average participant accuracy vs model size, we fit a power law of the form ax^{-b} . The 95% confidence intervals were estimated from the t-distribution of the sample mean.

Duration statistics: In the main text, we discussed the finding that the ability of human participants to distinguish model and human generated news articles decreases as our models become larger. We have also found that the average time spent for a given set of questions increases as the model size increases, as shown in Figure E.1. Lower

模型	参与者招募	参与者排除	性别 (m:f:other)	Mean Age	平均词数 (人类:模型)
控制	76	7	32:37:0	39	216:216
GPT-3 小型	80	7	41:31:1	40	216:188
GPT-3 中型	80	7	46:28:2	39	216:202
GPT-3 大型	81	24	46:28:2	37	216:200
GPT-3 XL	79	14	32:32:1	38	216:199
GPT-3 27亿	80	11	36:33:0	40	216:202
GPT-3 67亿	76	5	46:28:2	37	216:195
GPT-3 130亿	81	13	46:28:2	37	216:209
GPT-3 175B	80	9	42:29:0	37	216:216

表E.1: 实验参与者详情及文章长度，用于评估人类检测 ~ 200 词模型生成的新闻文章。因互联网检查失败，部分参与者被排除。

Average time spent trying to detect model generated news article



图E.1: 随着模型大小的增加，参与者花费更多时间尝试识别每篇新闻文章是否为机器生成。控制模型的持续时间用虚线表示。最佳拟合线是对数尺度上的线性模型，带有95%置信区间。

在每个实验中，一半的参与者被随机分配到测验A，另一半被随机分配到测验B。每个测验包含25篇文章：一半（12-13篇）是人类撰写的，另一半（12-13篇）是模型生成的：测验A中人类撰写的文章在测验B中对应模型生成的文章，反之亦然。每个参与者的测验问题顺序被打乱。参与者可以留言，并被要求说明是否之前见过这些文章。参与者被指示在测验过程中不要查阅文章或其内容，并在测验结束后被询问是否在测验过程中查阅了任何内容。

统计检验: 为了比较不同运行的平均值，我们对每个模型相对于控制模型进行了独立组双样本t检验。该检验在Python中使用`scipy.stats.ttest_ind`函数实现。在平均参与者准确率与模型大小关系图的回归线绘制中，我们拟合了形式为 ax^{-b} 的幂律。95%置信区间是从样本均值的t分布估计的。

持续时间统计: 在正文部分，我们讨论了人类参与者区分模型生成和人类生成的新闻文章的能力随着模型规模增大而下降的发现。我们还发现，对于给定的一组问题，平均花费的时间随着模型规模的增加而增加，如图E.1所示。较低

Model	Participants Recruited	Participants Excluded	Genders (m:f:other)	Mean Age	Average Word Count (human:model)
Control	79	17	32:37:0	39	569:464
GPT-3 175B	81	19	32:30:0	40	569:498

Table E.2: Participant details and article lengths for the experiments investigating human detection of ~ 500 word model generated news articles. Participants were excluded due to internet check fails.

accuracy scores despite increased time investment from participants supports the finding that larger models generate harder-to-distinguish news articles.

Preliminary investigation of ~ 500 word articles: We recruited 160 unique US-based participants to take part in 2 experiments through Positly (details are given in Table E.2). We randomly selected 12 Reuters world news articles from late 2019 and created a context for GPT-3 175B that consisted of a single Reuters article not in this set of 12. We then used the article titles and Reuters locations to generate completions from GPT-3 175B and the 160M control model from the previous experiments. These were used to create two 12-question quizzes per model, each consisting of half human written and half model generated articles. Comprehension questions were added and articles were shown to participants in 3 stages at 30 second intervals to encourage closer reading. Participants were paid \$12 for this task. Model generation selection methods, exclusion criteria, and statistical tests mirror those of the previous experiments.

F Additional Samples from GPT-3

GPT-3 adapts well to many tasks other than the ones explored in the main body of the paper. As an example, in Figure F.1, we show four uncurated samples from a prompt suggesting that the model write a poem, with a given title, in the style of Wallace Stevens. We first experimented with a few prompts, then generated four samples with no additional editing or selection (sampling at temperature 1 using nucleus sampling [HBFC19] with $P = 0.9$). Completions were truncated when the model began to write a new title and author heading, or broke into prose commentary.

模型	参与者招募	参与者排除	性别 (m:f:other)	Mean Age	平均词数 (人类:模型)
控制	79	17	32:37:0	39	569:464
GPT-3 175B	81	19	32:30:0	40	569:498

表 E.2: 用于研究人类检测 ~ 500 词模型生成的新闻文章的实验中参与者详情和文章长度。由于互联网检查失败，参与者被排除在外。

尽管参与者投入了更多时间，准确率得分仍然支持这一发现：较大的模型生成的新闻文章更难区分。

初步调查：~ 500 词文章：我们通过Positly招募了160名美国参与者参加2项实验（详细信息见表 E.2）。我们从2019年末随机选择了12篇路透社世界新闻文章，并为GPT-3 175B创建了一个上下文，该上下文包含一个不在这12篇文章集中的路透社文章。然后我们使用文章标题和路透社位置从GPT-3 175B和之前实验中的160M控制模型生成补全。这些被用于为每个模型创建两个12题的测验，每个测验包含一半人类撰写和一半模型生成的文章。添加了理解问题，并将文章以30秒的间隔向参与者展示三次，以鼓励仔细阅读。参与者为此任务获得12美元报酬。模型生成选择方法、排除标准和统计测试与之前的实验相同。

F 来自 GPT-3 的额外样本

GPT-3 能很好地适应论文主体部分探讨之外的多项任务。例如，在图F.1中，我们展示了四个来自提示的未筛选样本，提示要求模型以华莱士·史蒂文斯 (Wallace Stevens) 的风格写一首诗，并给定标题。我们首先尝试了几种提示，然后生成了四个样本，未进行任何额外的编辑或选择（使用温度为 1 的核采样 [HBFC 19]，并使用 $P = 0.9$ ）。当模型开始写新标题和作者署名时，补全内容会被截断，或者中断为散文式评论。

Context →	The City BY C. P. CAVAFY TRANSLATED BY EDMUND KEELEY [Poem text omitted]
	SOME TREES John Ashbery [Poem text omitted]
	Shadows on the Way Wallace Stevens
----- Generated Poem 1 -----	----- Generated Poem 3 -----
I must have shadows on the way If I am to walk I must have Each step taken slowly and alone To have it ready made And I must think in lines of grey To have dim thoughts to be my guide Must look on blue and green And never let my eye forget That color is my friend And purple must surround me too The yellow of the sun is no more Intrusive than the bluish snow That falls on all of us. I must have Grey thoughts and blue thoughts walk with me If I am to go away at all.	The sun was all we had. Now, in the shade All is changed. The mind must dwell on those White fields, that to its eyes were always old; Those ancient gleams, convoluting The way they lay among the huge roots, The great dark tomes of reverie, The plumed labyrinth of the sea. When our own breath gone faint and fade Up the long stairway of the dead We shall remember well The blindness of the light, the whiteness Of that white land. The footsteps, and The grasses in that light, and in the shade The menacing life. We played, too, once, in the warmest rooms. To one content, there is one content For all mankind. The forest glades Are all the more for being in fable, And wood and town in legend shadowed- Once more to return. The one content Has in it every happiness, more brave For being purposeless; and so Those island gardens, deep in the mind, Make of remoteness familiarity And of the unspoken name a friend.
----- Generated Poem 2 -----	----- Generated Poem 4 -----
There is a way in the world. That way a man may live in a small town His lifetime and see no man come, Never the stir of men about a market, Never the bobbing of covered carts, Never the creak passed on the stair Closing some other life; an old man Can prosper in that town, can prosper When his store is seen to close, his wife Fades in his house, and the world shouts in vain That no one comes to hear him Where his small bell tolls. In that small town Everything is known. The ways of the young men are known, And the ways of the young women As they walk to the Spring; The rumors are all known, The quarrels, the hurt, the envy Of the men and the women. Till a day comes That way out of the world That leads to a wood. Young, the young man Walks, a lover now, Knows the streets as they are and is of them. Comes to the wood. There, in the wood, among trees, He sees shadows on the way, hears voices, hears the wind and the rustling of leaves; Through an open glade He sees a shape and the shape hears: It waits as he waits, Waits as the shadows wait, As the voices wait; Shadows on the way, voices in the wind.	Nobody will come to this place. It is a road that leads nowhere. The solitude is deep. The mountains are high. But they are desolate, and they turn the traveler's face Towards the North. All the sounds of the world are far away. When the wind rises above the trees, The boughs bow to the ground. Even the birds that inhabit the tangle of weeds That is the roadside cover, are silent. One listens, But hears no roar of the forest. One is alone. One will be taken. One will be taken. There is no utterance, there is no conversation, But one is uneasy all the same.... There is a thin blue mist, A darkness rising like smoke, And within that darkness A possession of the heart. One will be taken.... It was here, and it will be here again- Here, under this sky empty and full of light.

Context →	The City BY C. P. CAVAFY TRANSLATED BY EDMUND KEELEY [Poem text omitted]
	SOME TREES John Ashbery [Poem text omitted]
	Shadows on the Way Wallace Stevens
----- Generated Poem 1 -----	----- Generated Poem 3 -----
I must have shadows on the way If I am to walk I must have Each step taken slowly and alone To have it ready made And I must think in lines of grey To have dim thoughts to be my guide Must look on blue and green And never let my eye forget That color is my friend And purple must surround me too The yellow of the sun is no more Intrusive than the bluish snow That falls on all of us. I must have Grey thoughts and blue thoughts walk with me If I am to go away at all.	The sun was all we had. Now, in the shade All is changed. The mind must dwell on those White fields, that to its eyes were always old; Those ancient gleams, convoluting The way they lay among the huge roots, The great dark tomes of reverie, The plumed labyrinth of the sea. When our own breath gone faint and fade Up the long stairway of the dead We shall remember well The blindness of the light, the whiteness Of that white land. The footsteps, and The grasses in that light, and in the shade The menacing life. We played, too, once, in the warmest rooms. To one content, there is one content For all mankind. The forest glades Are all the more for being in fable, And wood and town in legend shadowed- Once more to return. The one content Has in it every happiness, more brave For being purposeless; and so Those island gardens, deep in the mind, Make of remoteness familiarity And of the unspoken name a friend.
----- Generated Poem 2 -----	----- Generated Poem 4 -----
There is a way in the world. That way a man may live in a small town His lifetime and see no man come, Never the stir of men about a market, Never the bobbing of covered carts, Never the creak passed on the stair Closing some other life; an old man Can prosper in that town, can prosper When his store is seen to close, his wife Fades in his house, and the world shouts in vain That no one comes to hear him Where his small bell tolls. In that small town Everything is known. The ways of the young men are known, And the ways of the young women As they walk to the Spring; The rumors are all known, The quarrels, the hurt, the envy Of the men and the women. Till a day comes That way out of the world That leads to a wood. Young, the young man Walks, a lover now, Knows the streets as they are and is of them. Comes to the wood. There, in the wood, among trees, He sees shadows on the way, hears voices, hears the wind and the rustling of leaves; Through an open glade He sees a shape and the shape hears: It waits as he waits,Waits as the shadows wait, As the voices waitShadows on the way, voices in the wind.	Nobody will come to this place. It is a road that leads nowhere. The solitude is deep. The mountains are high. But they are desolate, and they turn the traveler's face Towards the North. All the sounds of the world are far away. When the wind rises above the trees, The boughs bow to the ground. Even the birds that inhabit the tangle of weeds That is the roadside cover, are silent. One listens, But hears no roar of the forest. One is alone. One will be taken. One will be taken. There is no utterance, there is no conversation, But one is uneasy all the same.... There is a thin blue mist, A darkness rising like smoke, And within that darkness A possession of the heart. One will be taken.... It was here, and it will be here again- Here, under this sky empty and full of light.

Figure F.1: Four uncurated completions from a context suggesting the model compose a poem in the style of Wallace Stevens with the title 'Shadows on the Way'.

图F.1: 四个未经筛选的补全结果, 来自一个提示模型以华莱士·史蒂文斯风格创作题为'路上的阴影'的诗的上下文。

G Details of Task Phrasing and Specifications

The following figures illustrate the formatting and phrasing of all the tasks included in the paper. All data comes from the ground truth datasets in this section, and no samples from GPT-3 are included here.

Context →	Article: Informal conversation is an important part of any business relationship. Before you start a discussion, however, make sure you understand which topics are suitable and which are considered taboo in a particular culture. Latin Americans enjoy sharing information about their local history, art and customs. You may expect questions about your family, and be sure to show pictures of your children. You may feel free to ask similar questions of your Latin American friends. The French think of conversation as an art form, and they enjoy the value of lively discussions as well as disagreements. For them, arguments can be interesting and they can cover pretty much or any topic ---- as long as they occur in a respectful and intelligent manner. In the United States, business people like to discuss a wide range of topics, including opinions about work, family, hobbies, and politics. In Japan, China, and Korea, however, people are much more private. They do not share much about their thoughts, feelings, or emotions because they feel that doing so might take away from the harmonious business relationship they're trying to build. Middle Easterners are also private about their personal lives and family matters. It is considered rude, for example, to ask a businessman from Saudi Arabia about his wife or children. As a general rule, it's best not to talk about politics or religion with your business friends. This can get you into trouble, even in the United States, where people hold different religious views. In addition, discussing one's salary is usually considered unsuitable. Sports is typically a friendly subject in most parts of the world, although be careful not to criticize national sport. Instead, be friendly and praise your host's team. Q: What shouldn't you do when talking about sports with colleagues from another country? A: Criticizing the sports of your colleagues' country. Q: Which is typically a friendly topic in most places according to the author? A: Sports. Q: Why are people from Asia more private in their conversation with others? A: They don't want to have their good relationship with others harmed by informal conversation. Q: The author considers politics and religion . . . A:
Correct Answer →	taboo
Incorrect Answer →	cheerful topics
Incorrect Answer →	rude topics
Incorrect Answer →	topics that can never be talked about

Figure G.1: Formatted dataset example for RACE-h. When predicting, we normalize by the unconditional probability of each answer as described in 2.

G 任务措辞和规范详解

以下图表说明了论文中所有任务包含的格式和措辞。所有数据均来自本节中的真实数据集，此处不包含来自GPT-3的样本。

上下文 →	文章: 非正式对话是任何业务的重要组成部分关系。但在开始讨论之前,请确保你理解在特定场合中哪些话题适合讨论,哪些被视为禁忌文化。拉丁美洲人喜欢分享他们当地的历史、艺术和习俗。你可能会被问到关于你的家庭的问题,以及一定要展示你孩子的照片。你可以自由地询问你拉丁美洲朋友类似的问题。法国人认为对话是一种艺术形式,他们喜欢生动讨论的价值,也喜欢活跃讨论的价值。 分歧。对他们来说,争论可以很有趣,而且可以涵盖几乎所有或任何话题——只要它们以尊重和智能的方式发生。 在美国,商人喜欢讨论广泛的话题,包括关于工作、家庭、爱好和政治的看法。 日本、中国和韩国,然而,人们却更加注重隐私。他们不会分享太多关于他们的想法、感受或情绪,因为他们觉得这样做可能会破坏他们试图建立的和谐商业关系。中东人同样对自己的私生活和家庭事务比较注重隐私。例如,向来自沙特阿拉伯的商人询问他的妻子或孩子被认为是不礼貌的。向来自沙特阿拉伯的商人询问他的妻子或孩子被认为是不礼貌的。通常情况下,最好不要和商业伙伴谈论政治或宗教。这可能会让你陷入麻烦,即使在宗教信仰各不相同的美利坚合众国也是如此。此外,讨论个人薪水通常被认为是不合适的。体育在世界上大多数地区通常是一个友好的话题,但要注意不要批评国家体育。相反,要友好地称赞东道主球队。相反,要友好地称赞东道主球队。 Q: 与来自另一个国家的同事谈论体育时,你不应该做什么?另一个国家? A: 批评同事国家的体育。 Q: 根据作者的说法,在大多数地方通常是一个友好的话题是什么?作者? A: 体育。 Q: 为什么亚洲人在与他人交谈时更注重隐私? A: 他们不希望自己的良好人际关系因非正式交谈而受到损害。非正式对话。 Q: 作者认为政治和宗教 . . . A:
正确答案 →	是禁忌。
错误答案 →	令人愉快的主题
错误答案 →	粗鲁的主题
错误答案 →	永远无法谈论的主题

图G.1: 为RACE-h准备的格式化数据集示例。在预测时,我们根据文中所述对每个答案的无条件概率进行归一化。2。

Context →	anli 2: anli 2: The Gold Coast Hotel & Casino is a hotel and casino located in Paradise, Nevada. This locals' casino is owned and operated by Boyd Gaming. The Gold Coast is located one mile (~ 1.6km) west of the Las Vegas Strip on West Flamingo Road. It is located across the street from the Palms Casino Resort and the Rio All Suite Hotel and Casino. Question: The Gold Coast is a budget-friendly casino. True, False, or Neither?
Correct Answer →	Neither
Incorrect Answer →	True
Incorrect Answer →	False

Figure G.2: Formatted dataset example for ANLI R2

Context →	Article: Mrs. Smith is an unusual teacher. Once she told each student to bring along a few potatoes in plastic bag. On each potato the students had to write a name of a person that they hated And the next day, every child brought some potatoes. Some had two potatoes;some three;some up to five. Mrs. Smith then told the children to carry the bags everywhere they went, even to the toilet, for two weeks. As day after day passed, the children started to complain about the awful smell of the rotten potatoes. Those children who brought five potatoes began to feel the weight trouble of the bags. After two weeks, the children were happy to hear that the game was finally ended. Mrs. Smith asked,"How did you feel while carrying the potatoes for two weeks?" The children started complaining about the trouble loudly. Then Mrs. Smith told them why she asked them to play the game. She said,"This is exactly the situation when you carry your hatred for somebody inside your heart. The terrible smell of the hatred will pollute your heart and you will carry something unnecessary with you all the time. If you cannot stand the smell of the rotten potatoes for just two weeks, can you imagine how heavy it would be to have the hatred in your heart for your lifetime? So throw away any hatred from your heart, and you'll be really happy." Q: Which of the following is True according to the passage? A: If a kid hated four people,he or she had to carry four potatoes. Q: We can learn from the passage that we should . . A: throw away the hatred inside Q: The children complained about _ besides the weight trouble. A: the smell Q: Mrs.Smith asked her students to write _ on the potatoes. A:
Correct Answer →	names
Incorrect Answer →	numbers
Incorrect Answer →	time
Incorrect Answer →	places

Figure G.3: Formatted dataset example for RACE-m. When predicting, we normalize by the unconditional probability of each answer as described in 2.

上下文 →	ANLI 2: ANLI 2: 金海岸酒店和赌场是一家酒店和赌场位于天堂, 内华达州。这家当地人的赌场由其所有和经营由鲍德游戏公司提供。黄金海岸位于一英里 (~) 处, 1.6公里) 的西弗拉明戈路上, 它位于街对面 位于拉斯维加斯大道西侧。 棕榈赌场度假村和里奥全套房酒店及赌场的对面。 问题: 黄金海岸是一个经济实惠的赌场。 都不是?
正确答案 →	都不是
错误答案 →	True
错误答案 →	错误

图G.2: ANLI R2的格式化数据集示例

上下文 →	文章: Mrs. 史密斯是一位不寻常的老师。她曾经告诉每个学生要带一些土豆在塑料袋里。在每个土豆上, 学生们必须写上他们讨厌的一个人的名字。第二天, 每个孩子都带来了一些土豆。有些人有两个土豆; 有些人三个; 有些人甚至多达五个。Mrs. 史密斯随后告诉孩子们, 无论他们去哪里都要带着袋子, 甚至去厕所, 持续两周。随着时间的推移, 孩子们开始抱怨那些腐烂土豆的难闻气味。那些带来五个土豆的孩子开始感到重量上的困扰。这些包的。两周后, 孩子们很高兴听说那个游戏终于结束了。Mrs. 史密斯问道: "你带着这些土豆两周时感觉怎么样?" 两周时, 孩子们开始大声抱怨这些麻烦。 然后夫人 史密斯告诉他们她为什么让他们玩游戏。 She 他 said,"这就是你带着对某人的仇恨时的情形" If 在你的心里。憎恨的可怕气味会污染你的心, 你将一直带着不必要的负担。 你连烂土豆的气味两周都受不了, 对吧 你想象一下, 如果心里一直怀着对{name}的仇恨, 那该有多沉重一辈子呢? 所以把心里的仇恨都扔掉吧, 你才能真正快乐。 。 Q: 根据文章, 以下哪项是正确的? A: 如果一个小孩子讨厌四个人, 他或她必须携带四个土豆。 Q: 我们可以从文中得知, 我们应该 . . A: 抛弃内心的仇恨 Q: 孩子们抱怨 _ 除了重量问题。 A: 气味 Q: 史密斯夫人让她学生们写 _ 在土豆上。 A:
正确答案 →	名称
错误答案 →	数字
错误答案 →	time
错误答案 →	地点

图G.3: 格式化 RACE-m的数据集示例。在预测时, 我们按每个答案的无条件概率进行归一化每个答案的描述 描述在 2。

Context →	How to apply sealant to wood.
Correct Answer →	Using a brush, brush on sealant onto wood until it is fully saturated with the sealant.
Incorrect Answer →	Using a brush, drip on sealant onto wood until it is fully saturated with the sealant.

Figure G.4: Formatted dataset example for PIQA

Context →	My body cast a shadow over the grass because
Correct Answer →	the sun was rising.
Incorrect Answer →	the grass was cut.

Figure G.5: Formatted dataset example for COPA

Context →	(CNN) Yuval Rabin, whose father, Yitzhak Rabin, was assassinated while serving as Prime Minister of Israel, criticized Donald Trump for appealing to "Second Amendment people" in a speech and warned that the words that politicians use can incite violence and undermine democracy. "Trump's words are an incitement to the type of political violence that touched me personally," Rabin wrote in USA Today. He said that Trump's appeal to "Second Amendment people" to stop Hillary Clinton -- comments that were criticized as a call for violence against Clinton, something Trump denied -- "were a new level of ugliness in an ugly campaign season." - The son of a former Israeli Prime Minister who was assassinated wrote an op ed about the consequence of violent political rhetoric. - Warns of "parallels" between Israel of the 1990s and the U.S. today.
Correct Answer →	- Referencing his father, who was shot and killed by an extremist amid political tension in Israel in 1995, Rabin condemned Donald Trump's aggressive rhetoric.
Correct Answer →	- Referencing his father, who was shot and killed by an extremist amid political tension in Israel in 1995, Rabin condemned Trump's aggressive rhetoric.
Incorrect Answer →	- Referencing his father, who was shot and killed by an extremist amid political tension in Israel in 1995, Rabin condemned Hillary Clinton's aggressive rhetoric.
Incorrect Answer →	- Referencing his father, who was shot and killed by an extremist amid political tension in Israel in 1995, Rabin condemned U.S.'s aggressive rhetoric.
Incorrect Answer →	- Referencing his father, who was shot and killed by an extremist amid political tension in Israel in 1995, Rabin condemned Yitzhak Rabin's aggressive rhetoric.

Figure G.6: Formatted dataset example for ReCoRD. We consider the context above to be a single "problem" because this is how the task is presented in the ReCoRD dataset and scored in the ReCoRD evaluation script.

Context →	anli 1: anli 1: Fulton James MacGregor MSP is a Scottish politician who is a Scottish National Party (SNP) Member of Scottish Parliament for the constituency of Coatbridge and Chryston. MacGregor is currently Parliamentary Liaison Officer to Shona Robison, Cabinet Secretary for Health & Sport. He also serves on the Justice and Education & Skills committees in the Scottish Parliament. Question: Fulton James MacGregor is a Scottish politician who is a Liaison officer to Shona Robison who he swears is his best friend. True, False, or Neither?
Correct Answer →	Neither
Incorrect Answer →	True
Incorrect Answer →	False

Figure G.7: Formatted dataset example for ANLI R1

上下文 →	如何将密封胶涂抹在木材上。
正确答案 →	使用刷子，将密封胶涂抹在木材上，直到木材完全浸透密封胶。
错误答案 →	用刷子蘸上密封胶滴在木头上，直到木头完全浸透密封胶。

图G.4: PIQA的格式化数据集示例

上下文 →	我的身体在草地上投下了影子，因为
正确答案 →	太阳正在升起。
错误答案 →	草被割了。

图G.5: COPA的格式化数据集示例

上下文 →	(CNN) 尤瓦尔·拉宾，其父亲，伊扎克·拉宾，被暗杀时，曾批评唐纳德·特朗普呼吁"第二修正案支持者"。在一次演讲中，并警告称政客们使用的言辞可能煽动暴力、破坏民主。"特朗普的" words是煽动那种政治暴力的言辞，这种暴力触及了"我本人，"拉宾在《USA Today》上写道。他说特朗普的呼吁"第二修正案支持者"阻止希拉里·克林顿——评论称批评为呼吁对克林顿施加暴力，特朗普否认了这一点 -- "成为了丑陋竞选季节中一个更糟糕的新水平。"
正确答案 →	- 提及他的父亲，后者被一名极端主义者枪杀身亡 1995年以色列的政治紧张局势，拉宾谴责了唐纳德·特朗普的激烈的言辞。
正确答案 →	- 提及他的父亲，在1995年以色列政治紧张期间被极端分子枪杀身亡，拉宾谴责了特朗普的激烈言辞。
错误答案 →	- 提及他的父亲，后者被一名极端主义者枪杀身亡 政治紧张时期在以色列，拉宾谴责了希拉里·克林顿的强硬言论。
错误答案 →	- 提及他的父亲，这位在1995年以色列政治紧张时期被极端分子枪杀身亡的人，拉宾谴责了美国的强硬言论。
错误答案 →	- 提及他的父亲，后者被一名极端主义者枪杀身亡 1995年以色列的政治紧张局势中，拉宾谴责了伊扎克·拉宾攻击性言辞。

图G.6: ReCoRD的格式化数据集示例。我们认为上述上下文是一个单一的“问题”，因为这是ReCoRD数据集中任务呈现的方式，并在ReCoRD评估脚本中进行评分。

上下文 →	ANLI 1: ANLI 1: Fulton James MacGregor MSP 是一位苏格兰政治家他是苏格兰民族党 (SNP) 的苏格兰议会成员 MacGregor 目前是代表 Coatbridge and Chryston 选区。沙·罗宾森议长 (苏格兰议会议长) 的议会联络官，负责健康与体育事务。他还担任苏格兰议会的司法与教育与技术委员会成员。 问题: Fulton James MacGregor 是一位苏格兰政治家，担任联络官他是 Shona Robison 的联络官，他发誓她是他最好的朋友 True, 错误, or 都不是?
正确答案 →	都不是
错误答案 →	True
错误答案 →	错误

图G.7: ANLI R1的格式化数据集示例

Context →	Organisms require energy in order to do what?
Correct Answer →	mature and develop.
Incorrect Answer →	rest soundly.
Incorrect Answer →	absorb light.
Incorrect Answer →	take in nutrients.

Figure G.8: Formatted dataset example for OpenBookQA. When predicting, we normalize by the unconditional probability of each answer as described in 2.

Context →	Making a cake: Several cake pops are shown on a display. A woman and girl are shown making the cake pops in a kitchen. They
Correct Answer →	bake them, then frost and decorate.
Incorrect Answer →	taste them as they place them on plates.
Incorrect Answer →	put the frosting on the cake as they pan it.
Incorrect Answer →	come out and begin decorating the cake as well.

Figure G.9: Formatted dataset example for HellaSwag

Context →	anli 3: anli 3: We shut the loophole which has American workers actually subsidizing the loss of their own job. They just passed an expansion of that loophole in the last few days: \$43 billion of giveaways, including favors to the oil and gas industry and the people importing ceiling fans from China. Question: The loophole is now gone True, False, or Neither?
Correct Answer →	False
Incorrect Answer →	True
Incorrect Answer →	Neither

Figure G.10: Formatted dataset example for ANLI R3

Context →	Question: George wants to warm his hands quickly by rubbing them. Which skin surface will produce the most heat? Answer:
Correct Answer →	dry palms
Incorrect Answer →	wet palms
Incorrect Answer →	palms covered with oil
Incorrect Answer →	palms covered with lotion

Figure G.11: Formatted dataset example for ARC (Challenge). When predicting, we normalize by the unconditional probability of each answer as described in 2.

Context →	lull is to trust as
Correct Answer →	cajole is to compliance
Incorrect Answer →	balk is to fortitude
Incorrect Answer →	betray is to loyalty
Incorrect Answer →	hinder is to destination
Incorrect Answer →	soothe is to passion

Figure G.12: Formatted dataset example for SAT Analogies

Correct Context →	Grace was happy to trade me her sweater for my jacket. She thinks the sweater
Incorrect Context →	Grace was happy to trade me her sweater for my jacket. She thinks the jacket
Target Completion →	looks dowdy on her.

Figure G.13: Formatted dataset example for Winograd. The ‘partial’ evaluation method we use compares the probability of the completion given a correct and incorrect context.

背景 →	生物需要能量才能做什么?
正确答案 →	成熟和发展。
错误答案 →	好好休息。
错误答案 →	吸收光。
错误答案 →	吸收营养。

图G.8: OpenBookQA的格式化数据集示例。预测时，我们按每个答案的无条件概率进行归一化，如2所述。

上下文 →	制作蛋糕: 展示台上摆放着几个蛋糕棒。 一位妇女和一位女孩正在厨房里制作蛋糕棒。 They
正确答案 →	烘烤它们，然后结霜并装饰。
错误答案 →	品尝它们放置在盘子上的样子。
错误答案 →	在烘烤蛋糕时将糖霜放在蛋糕上。
错误答案 →	也出来开始装饰蛋糕。

图G.9: HellaSwag的格式化数据集示例

上下文 →	ANLI 3: ANLI 3: 我们堵上了让美国工人实际上补贴他们自己失业的损失。 他们刚通过了一项该漏洞的扩张，就在最近几天: 430亿美元的福利，包括石油和天然气行业的优惠，以及从中国进口吊扇的人们的优惠。 问题: 漏洞现已消失。对，错，还是都不对?
正确答案 →	错
错误答案 →	True
错误答案 →	既非

图G.10: ANLI R3的格式化数据集示例

上下文 →	问题: 乔治想通过搓手快速暖和双手。 哪个皮肤表面会产生最多热量? 答案:
正确答案 →	干燥的手掌
错误答案 →	湿润的手掌
错误答案 →	手掌上沾满油
错误答案 →	手掌上涂着润肤露

图G.11: ARC (挑战) 的格式化数据集示例。预测时，我们按照文中所述对每个答案的无条件概率进行归一化2。

上下文 →	哄骗是信任
正确答案 →	说服是服从
错误答案 →	犹豫是勇气
错误答案 →	背叛是对忠诚的
错误答案 →	阻碍是对目的地的
错误答案 →	抚慰是对激情的

图G.12: 为SAT类比的格式化数据集示例

正确语境 →	恩典很高兴用她的毛衣和我交换夹克。 她认为
错误语境 →	这件夹克 恩典很高兴用她的毛衣和我交换夹克。 她觉得
目标补全 →	穿在她身上显得很没精神。

图G.13: Winograd的格式化数据集示例。我们使用的‘部分’评估方法比较了在正确和错误上下文中给出的补全概率。

Correct Context →	Johnny likes fruits more than vegetables in his new keto diet because the fruits
Incorrect Context →	Johnny likes fruits more than vegetables in his new keto diet because the vegetables
Target Completion →	are saccharine.

Figure G.14: Formatted dataset example for Winogrande. The ‘partial’ evaluation method we use compares the probability of the completion given a correct and incorrect context.

Context →	<p>READING COMPREHENSION ANSWER KEY</p> <p>While this process moved along, diplomacy continued its rounds. Direct pressure on the Taliban had proved unsuccessful. As one NSC staff note put it, "Under the Taliban, Afghanistan is not so much a state sponsor of terrorism as it is a state sponsored by terrorists." In early 2000, the United States began a high-level effort to persuade Pakistan to use its influence over the Taliban. In January 2000, Assistant Secretary of State Karl Inderfurth and the State Department’s counterterrorism coordinator, Michael Sheehan, met with General Musharraf in Islamabad, dangling before him the possibility of a presidential visit in March as a reward for Pakistani cooperation. Such a visit was coveted by Musharraf, partly as a sign of his government’s legitimacy. He told the two envoys that he would meet with Mullah Omar and press him on Bin Laden. They left, however, reporting to Washington that Pakistan was unlikely in fact to do anything," given what it sees as the benefits of Taliban control of Afghanistan." President Clinton was scheduled to travel to India. The State Department felt that he should not visit India without also visiting Pakistan. The Secret Service and the CIA, however, warned in the strongest terms that visiting Pakistan would risk the President’s life. Counterterrorism officials also argued that Pakistan had not done enough to merit a presidential visit. But President Clinton insisted on including Pakistan in the itinerary for his trip to South Asia. His one-day stopover on March 25, 2000, was the first time a U.S. president had been there since 1969. At his meeting with Musharraf and others, President Clinton concentrated on tensions between Pakistan and India and the dangers of nuclear proliferation, but also discussed Bin Laden. President Clinton told us that when he pulled Musharraf aside for a brief, one-on-one meeting, he pleaded with the general for help regarding Bin Laden." I offered him the moon when I went to see him, in terms of better relations with the United States, if he’d help us get Bin Laden and deal with another issue or two." The U.S. effort continued.</p> <p>Who did The State Department feel should visit both India and Pakistan?</p>
Correct Answer →	- [False] Bin Laden
Incorrect Answer →	- [True] Bin Laden

Figure G.15: Formatted dataset example for MultiRC. There are three levels within MultiRC: (1) the passage, (2) the questions, and (3) the answers. During evaluation, accuracy is determined at the per-question level, with a question being considered correct if and only if all the answers within the question are labeled correctly. For this reason, we use K to refer to the number of **questions** shown within the context.

Context →	<p>Question: Which factor will most likely cause a person to develop a fever?</p> <p>Answer:</p>
Correct Answer →	a bacterial population in the bloodstream
Incorrect Answer →	a leg muscle relaxing after exercise
Incorrect Answer →	several viral particles on the skin
Incorrect Answer →	carbohydrates being digested in the stomach

Figure G.16: Formatted dataset example for ARC (Easy). When predicting, we normalize by the unconditional probability of each answer as described in 2.

正确上下文 →	约翰尼在他的新生酮饮食中更喜欢水果而不是蔬菜，因为水果
错误上下文 →	约翰尼在他的新生酮饮食中更喜欢水果而不是蔬菜，因为蔬菜
目标补全 →	是甜腻的。

图G.14: Winogrande的格式化数据集示例。我们使用的‘部分’评估方法比较了在正确和错误上下文中给出的补全的概率。

上下文 →	<p>阅读理解答案键虽然这个过程在进行中，外交仍在继续。对塔利班的直接压力已被证明是无效的。正如一位NSC工作人员的笔记所说，‘在塔利班统治下，阿富汗与其说是恐怖主义的国家赞助者，不如说是由恐怖分子赞助的国家。’2000年初，美国开始了一项高级别的工作，试图说服巴基斯坦利用其对塔利班的影响力。2000年1月，美国国务院负责反恐的助理国务卿卡尔·英德弗思和迈克尔·希恩与穆沙拉夫将军在伊斯兰堡会面，向他暗示3月份有可能进行总统访问，作为对巴基斯坦合作的奖励。这样的访问被穆沙拉夫渴望，部分是因为它是其政府合法性的象征。他告诉两位使节，他会与奥马尔·哈基姆会面，并对他施加压力，让他交出本·拉登。然而，他们离开时向华盛顿报告说，巴基斯坦实际上不太可能采取任何行动，‘考虑到塔利班控制阿富汗的所谓好处。’克林顿总统原定前往印度。国务院认为，他不应该在没有访问巴基斯坦的情况下访问印度。然而，美国特勤局和中央情报局以最强烈的言辞警告说，访问巴基斯坦将危及总统的生命。反恐官员还认为，巴基斯坦没有做到足以配得上总统访问的程度。但克林顿总统坚持要求将巴基斯坦纳入他南亚之行的行程中。2000年3月25日的一日停留是自1969年以来美国总统首次到访。在会晤穆沙拉夫和其他人的时候，克林顿总统专注于巴基斯坦和印度之间的紧张关系以及核扩散的危险，但也讨论了本·拉登。克林顿总统告诉我们，当他把穆沙拉夫拉到一旁进行简短的一对一会面时，他恳求这位将军帮助我们处理本·拉登的问题。“我去看他时向他提供了月亮，如果他能帮助我们抓住本·拉登并处理其他几个问题，我们就会与美国建立更好的关系。”美国的工作仍在继续。</p> <p>美国国务院认为谁应该访问印度和巴基斯坦？</p>
正确答案 →	- [错误] 本·拉登
错误答案 →	- [正确] 本·拉登

图G.15: MultiRC的格式化数据集示例。MultiRC中有三个层级：（1）段落，（2）问题，（3）答案。在评估过程中，准确率是在每个问题级别确定的，只有当问题中的所有答案都被正确标记时，该问题才被视为正确。因此，我们使用 K 来指代上下文中显示的问题数量。

上下文 →	<p>问题: 哪种因素最可能导致一个人发烧?</p> <p>答案:</p>
正确答案 →	血液中的细菌种群
错误答案 →	运动后腿部肌肉放松
错误答案 →	皮肤上的多个病毒颗粒
错误答案 →	碳水化合物在胃中消化

图G.16: ARC (简单) 的格式化数据集示例。预测时，我们按照文中所述对每个答案的无条件概率进行归一化2。

Context →	Bob went to the gas station to fill up his car. His tank was completely empty and so was his wallet. The cashier offered to pay for his gas if he came back later to pay. Bob felt grateful as he drove home.
Correct Answer →	Bob believed that there were good people in the world.
Incorrect Answer →	Bob contemplated how unfriendly the world was.

Figure G.17: Formatted dataset example for StoryCloze

Context →	<p>Helsinki is the capital and largest city of Finland. It is in the region of Uusimaa, in southern Finland, on the shore of the Gulf of Finland. Helsinki has a population of , an urban population of , and a metropolitan population of over 1.4 million, making it the most populous municipality and urban area in Finland. Helsinki is some north of Tallinn, Estonia, east of Stockholm, Sweden, and west of Saint Petersburg, Russia. Helsinki has close historical connections with these three cities.</p> <p>The Helsinki metropolitan area includes the urban core of Helsinki, Espoo, Vantaa, Kauniainen, and surrounding commuter towns. It is the world's northernmost metro area of over one million people, and the city is the northernmost capital of an EU member state. The Helsinki metropolitan area is the third largest metropolitan area in the Nordic countries after Stockholm and Copenhagen, and the City of Helsinki is the third largest after Stockholm and Oslo. Helsinki is Finland's major political, educational, financial, cultural, and research center as well as one of northern Europe's major cities. Approximately 75% of foreign companies that operate in Finland have settled in the Helsinki region. The nearby municipality of Vantaa is the location of Helsinki Airport, with frequent service to various destinations in Europe and Asia.</p> <p>Q: what is the most populous municipality in Finland?</p> <p>A: Helsinki</p> <p>Q: how many people live there?</p> <p>A: 1.4 million in the metropolitan area</p> <p>Q: what percent of the foreign companies that operate in Finland are in Helsinki?</p> <p>A: 75%</p> <p>Q: what towns are a part of the metropolitan area?</p> <p>A:</p>
Target Completion →	Helsinki, Espoo, Vantaa, Kauniainen, and surrounding commuter towns

Figure G.18: Formatted dataset example for CoQA

Context →	Please unscramble the letters into a word, and write that word: asinoc =
Target Completion →	casino

Figure G.19: Formatted dataset example for Cycled Letters

上下文 →	鲍勃去加油站给他的车加满油。他的油箱完全是空的，他的钱包也是空的。收银员提出如果他去得晚些时候再付钱，可以帮他支付加油费。他可以回来支付。鲍勃开车回家时感到很感激。
正确答案 →	鲍勃相信世界上有好人。
错误答案 →	鲍勃思考这个世界多么不友好。

图G.17: StoryCloze的格式化数据集示例

上下文 →	<p>赫尔辛基是芬兰的首都和最大城市。它位于乌西马地区，芬兰南部，芬兰湾沿岸。赫尔辛基有人口，城市人口为，大都市超过140万人，使其成为芬兰人口最多、面积最大的市镇和城市地区。赫尔辛基位于塔林以北，爱沙尼亚，瑞典斯德哥尔摩以东，俄罗斯圣彼得堡以西。与这三个城市有着密切的历史联系。</p> <p>赫尔辛基大都会区包括赫尔辛基、埃斯波、万塔、库奥尼ainen以及周边的通勤城镇。它是世界上人口超过一百万的最北端大都市区，这座城市是欧盟成员国中最北的首都。赫尔辛基都市地区是北欧第三大都市区，仅次于斯德哥尔摩和哥本哈根，赫尔辛基市是第三大，仅次于斯德哥尔摩和奥斯陆。赫尔辛基是芬兰的主要政治、教育、金融、文化，以及研究中心，也是北欧主要城市之一。大约75%的外国公司在芬兰运营的那些机构已经定居在赫尔辛基地区。附近的万塔市是赫尔辛基机场的所在地，该机场有频繁的航班飞往欧洲和亚洲的各个目的地。服务至欧洲和亚洲的各个目的地。</p> <p>Q: 芬兰人口最多的市镇是哪个?</p> <p>A: 赫尔辛基</p> <p>Q: 那里有多少人居住?</p> <p>A: 大都会地区有140万人。</p> <p>Q: 在芬兰运营的外国公司中，有多少百分比位于赫尔辛基?</p> <p>A: 75%。</p> <p>Q: 哪些城镇属于大都会地区?</p> <p>A:</p>
目标补全 →	赫尔辛基、埃斯波、万塔、库奥尼ainen以及周边通勤城镇

图G.18: CoQA的格式化数据集示例

上下文 →	请将字母重组成一个单词，并写出那个单词: asinoc =
目标补全 →	casino

图G.19: 循环字母的格式化数据集示例

Context → Passage: Saint Jean de Brébeuf was a French Jesuit missionary who travelled to New France in 1625. There he worked primarily with the Huron for the rest of his life, except for a few years in France from 1629 to 1633. He learned their language and culture, writing extensively about each to aid other missionaries. In 1649, Brébeuf and another missionary were captured when an Iroquois raid took over a Huron village . Together with Huron captives, the missionaries were ritually tortured and killed on March 16, 1649. Brébeuf was beatified in 1925 and among eight Jesuit missionaries canonized as saints in the Roman Catholic Church in 1930. Question: How many years did Saint Jean de Brébeuf stay in New France before he went back to France for a few years? Answer:

Target Completion → 4

Figure G.20: Formatted dataset example for DROP

Context → Fill in blank:

She held the torch in front of her.

She caught her breath.

"Chris? There's a step."

"What?"

"A step. Cut in the rock. About fifty feet ahead." She moved faster. They both moved faster. "In fact," she said, raising the torch higher, "there's more than a ----. ->

Target Completion → step

Figure G.21: Formatted dataset example for LAMBADA

Context → Please unscramble the letters into a word, and write that word: skicts =

Target Completion → sticks

Figure G.22: Formatted dataset example for Anagrams 1 (A1)

Context → Please unscramble the letters into a word, and write that word: volwskagen =

Target Completion → volkswagen

Figure G.23: Formatted dataset example for Anagrams 2

Context → Q: Who played tess on touched by an angel?

A:

Target Completion → Delloreese Patricia Early (July 6, 1931 { November 19, 2017), known professionally as Della Reese

Figure G.24: Formatted dataset example for Natural Questions

背景 → 段落: 圣让·德·布勒贝夫是一位法国耶稣会传教士, 1625年前往新法兰西。在那里, 他主要与休伦人合作, 直到他生命的其余时间, 除了1629年到法国的几年外, 1633。他学习了他们的语言和文化, 广泛地写下了每一点, 以帮助其他传教士。在1649年, Brébeuf和另一位传教士一起在一场易洛魁袭击中俘虏, 当时袭击者占领了一个胡伦村庄。在Huron人质面前, 传教士们被举行仪式地折磨并杀害1649年3月16日。Brébeuf于1925年被宣福, 在八名耶稣会传教士于1930年被罗马天主教会封圣。问题: 圣让·德·布勒贝夫在新法兰西停留了多少年在他回到法国的几年之前? 回答:

目标补全 → 4

图G.20: DROP的格式化数据集示例

上下文 → 填空:

她手里举着火把。

她屏住了呼吸。

"克里斯?" "有个台阶。"

"什么?"

"一个台阶。" 切割岩石。前方大约五十英尺。"她移动得更快了。他们都移动得更快了。"事实上,"她说, 把火把举得更高。"有更多 ----. ->

Target 补全 → 步

图G.21: 为LAMBADA的格式化数据集示例

上下文 → 请将字母重组成一个单词, 并写出这个单词: skicts =

目标补全 → sticks

图G.22: 字母重组1 (A1) 格式化数据集示例

上下文 → 请将字母重组成一个单词, 并写出这个单词: volwskagen =

目标补全 → volkswagen

图G.23: 字母重组2的格式化数据集示例

上下文 → Q: 谁在《与天使有约》中扮演了Tess?

A:

目标补全 → 德洛里丝·帕特里夏·厄尔利 (1931年7月6日 { 2017年11月19日), 以德拉·里斯的专业身份为人所知

图G.24: 自然问题格式化数据集示例

Context → TITLE: William Perry (American football) - Professional career
 PARAGRAPH: In 1985, he was selected in the first round of the 1985 NFL Draft by the Chicago Bears; he had been hand-picked by coach Mike Ditka. However, defensive coordinator Buddy Ryan, who had a highly acrimonious relationship with Ditka, called Perry a "wasted draft-pick". Perry soon became a pawn in the political power struggle between Ditka and Ryan. Perry's "Refrigerator" nickname followed him into the NFL and he quickly became a favorite of the Chicago Bears fans. Teammates called him "Biscuit," as in "one biscuit shy of 350 pounds." While Ryan refused to play Perry, Ditka decided to use Perry as a fullback when the team was near the opponents' goal line or in fourth and short situations, either as a ball carrier or a lead blocker for star running back Walter Payton. Ditka stated the inspiration for using Perry as a fullback came to him during five-yard sprint exercises. During his rookie season, Perry rushed for two touchdowns and caught a pass for one. Perry even had the opportunity to run the ball during Super Bowl XX, as a nod to his popularity and contributions to the team's success. The first time he got the ball, he was tackled for a one-yard loss while attempting to throw his first NFL pass on a halfback option play. The second time he got the ball, he scored a touchdown (running over Patriots linebacker Larry McGrew in the process). About halfway through his rookie season, Ryan finally began to play Perry, who soon proved that he was a capable defensive lineman. His Super Bowl ring size is the largest of any professional football player in the history of the event. His ring size is 25, while the ring size for the average adult male is between 10 and 12. Perry went on to play for ten years in the NFL, retiring after the 1994 season. In his ten years as a pro, he regularly struggled with his weight, which hampered his performance at times. He played in 138 games, recording 29.5 sacks and five fumble recoveries, which he returned for a total of 71 yards. In his offensive career he ran five yards for two touchdowns, and had one reception for another touchdown. Perry later attempted a comeback, playing an unremarkable 1996 season with the London Monarchs of the World League of American Football (later NFL Europa).

Q: what team did he play for?

A:

Target Completion → the Chicago Bears

Figure G.25: Formatted dataset example for QuAC

Context → Please unscramble the letters into a word, and write that word:
 r e ! c . i p r o . c a / l =

Target Completion → reciprocal

Figure G.26: Formatted dataset example for Symbol Insertion

Context → Please unscramble the letters into a word, and write that word:
 taefed =

Target Completion → defeat

Figure G.27: Formatted dataset example for Reversed Words

背景 → 标题: 威廉·佩里 (美式橄榄球) - 职业生涯 段落: 1985年, 他被选入1985年NFL选秀首轮; 他曾被教练迈克·迪特卡亲自挑选。然而, 曾与迪特卡关系紧张且充满敌意的防守协调员巴迪·莱恩称佩里是“浪费的选秀顺位”。佩里很快成为迪特卡和莱恩之间政治权力斗争的棋子。佩里的“冰箱”绰号伴随他进入NFL, 他迅速成为芝加哥熊队球迷的宠儿。队友们称他为“饼干”, 意为“离350磅差一个饼干”。莱恩拒绝上场让佩里, 而迪特卡决定在球队接近对手端区或处于第四档短距离情况下使用佩里, 无论是作为持球跑卫还是明星跑卫瓦尔特·派顿的引导挡卫。迪特卡表示, 使用佩里作为全卫的灵感来自他在五码冲刺训练中的体验。在他新秀赛季, 佩里冲入两个达阵, 并接住一个传球得一分。佩里甚至有机会在超级碗XX中持球, 这是对他人气和为球队成功做出的贡献的认可。他第一次持球时, 在半卫选项进攻中尝试首次投出NFL传球时被擒获, 损失一码。第二次持球时, 他得分达阵 (过程中撞过爱国者队线卫拉里·麦格鲁)。在他新秀赛季过半时, 莱恩终于开始让佩里上场, 佩里很快证明了他是一名有能力的防守线卫。他的超级碗戒指尺寸是历史上任何职业橄榄球运动员中最大的。他的戒指尺寸为25号, 而普通成年男性的戒指尺寸在10至12号之间。佩里在NFL征战了十年, 1994赛季后退役。在他十年的职业生涯中, 他经常与体重问题作斗争, 这有时会影响到他的表现。他参加了138场比赛, 记录了29.5次擒抱和5次抄截, 抄截的回攻总得分为71码。在他的进攻生涯中, 他冲入两个达阵, 并有一次接球得一分。佩里后来尝试复出, 在1996赛季为世界橄榄球联盟 (后为NFL欧洲区) 的伦敦君主队打了一季并不起眼的比赛。问: 他效力于哪个球队? 答:

目标C 完成 → 芝加哥熊队

图G.25: QuAC的格式化数据集示例

上下文 → 请将字母重新排列成一个单词, 并写出那个单词:
 r e ! c . i p r o . c a / l =

Target 补全 → 互惠

图 G.26: 符号入的格式化数据集示例 `<code>sertion</code>`

上下文 → 请将字母重新排列成一个单词, 并写出这个单词:
 taefed =

目标补全 → defeat

图G.27: 用于反转单词的格式化数据集示例

Context →	Title: The Blitz
	Background: From the German point of view, March 1941 saw an improvement. The Luftwaffe flew 4,000 sorties that month, including 12 major and three heavy attacks. The electronic war intensified but the Luftwaffe flew major inland missions only on moonlit nights. Ports were easier to find and made better targets. To confuse the British, radio silence was observed until the bombs fell. X- and Y-Gerät beams were placed over false targets and switched only at the last minute. Rapid frequency changes were introduced for X-Gerät, whose wider band of frequencies and greater tactical flexibility ensured it remained effective at a time when British selective jamming was degrading the effectiveness of Y-Gerät.
	Q: How many sorties were flown in March 1941?
	A: 4,000
	Q: When did the Luftwaffe fly inland missions?
	A: only on moonlit nights
Target Completion →	only on moonlit nights

Figure G.28: Formatted dataset example for SQuADv2

Context →	Normal force -- In a simple case such as an object resting upon a table, the normal force on the object is equal but in opposite direction to the gravitational force applied on the object (or the weight of the object), that is, $N = m g$ ($\displaystyle N=mg$), where m is mass, and g is the gravitational field strength (about 9.81 m/s on Earth). The normal force here represents the force applied by the table against the object that prevents it from sinking through the table and requires that the table is sturdy enough to deliver this normal force without breaking. However, it is easy to assume that the normal force and weight are action-reaction force pairs (a common mistake). In this case, the normal force and weight need to be equal in magnitude to explain why there is no upward acceleration of the object. For example, a ball that bounces upwards accelerates upwards because the normal force acting on the ball is larger in magnitude than the weight of the ball.
	question: is the normal force equal to the force of gravity?
	answer: yes
Target Completion →	yes

Figure G.29: Formatted dataset example for BoolQ

Context →	The trend toward lower rents may seem surprising given that some communities in New York are bemoaning the loss of favorite local businesses to high rents. But, despite the recent softening, for many of these retailers there's still been too big a jump from the rental rates of the late 1970s, when their leases were signed. Certainly, the recent drop in prices doesn't mean Manhattan comes cheap.
	question: Manhattan comes cheap. true, false, or neither?
	answer: false
Target Completion →	false

Figure G.30: Formatted dataset example for CB

背景 →	标题: 闪电战
	背景: 从德军的角度来看, 1941年3月情况有所改善。当月, 空军执行了4,000次出击任务, 包括12次重大任务和三次重击。电子战加剧, 但德国空军只在月光下的夜晚执行主要的内陆任务。港口更容易被找到, 也更适合作为目标。为了迷惑英国人, 无线电静默被持续观察, 直到炸弹落下。X-和Y-Gerät光束被放置在假目标上, 并且只在最后一刻切换。快速频率的变化被引入X-Gerät, 其更宽的频率范围和更大的战术灵活性确保它在当时仍然有效。英国的干扰措施正在降低 Y-Gerät 的有效性。
	Q: 1941年3月飞行的出击次数是多少?
	A: 4,000
	Q: 德国空军何时飞越内陆任务?
	A: 仅在月光照耀的夜晚
目标: 补全 →	仅在月光照耀的夜晚

图G.28: 为SQuADv2准备的格式化数据集示例

Context →	法向力 —— 在一个简单的案例中, 例如一个物体放置在桌子上, 物体所受的正常力大小相等但方向相反。重力作用于物体 (或物体的重量), 即, $N = m g$ ($\displaystyle N=mg$), 其中 m 是质量, g 是重力场强度 (在地球上约为9.81米/秒)。此处表示桌子施加在物体上的力防止它穿过桌子并要求桌子足够坚固以承受这种支持力而不损坏。但是, 它很容易假设支持力和重力是作用-反作用力。很容易假设支持力和重力是作用-反作用力, 强制配对 (一个常见的错误)。在这种情况下, 法向力和重力需要大小相等来解释为什么物体没有向上的加速度。例如, 一个向上弹跳的球因为作用在球上的法向力更大而向上加速。在大小上比球的重量要大。
	问题: 法向力是否等于重力?
	答案: yes
目标补全 →	yes

图G.29: BoolQ的格式化数据集示例

上下文 →	降低租金的趋势可能看起来令人惊讶, 因为有些纽约的一些社区正为高租金导致喜爱的本地企业流失而哀叹。但是, 尽管最近有所缓和, 对许多这些零售商来说, 租金仍然跳涨得太厉害。在20世纪70年代末, 当时他们的租约签署了。当然, 最近的价格下降并不意味着曼哈顿变得便宜。
	问题: 曼哈顿变得便宜。是, 否, 还是都不是?
	回答: false
目标补全 →	false

图G.30: CB的格式化数据集示例

Context →	The bet, which won him dinner for four, was regarding the existence and mass of the top quark, an elementary particle discovered in 1995. question: The Top Quark is the last of six flavors of quarks predicted by the standard model theory of particle physics. True or False? answer:
Target Completion →	False

Figure G.31: Formatted dataset example for RTE

Context →	An outfitter provided everything needed for the safari. Before his first walking holiday, he went to a specialist outfitter to buy some boots. question: Is the word 'outfitter' used in the same way in the two sentences above? answer:
Target Completion →	no

Figure G.32: Formatted dataset example for WiC

Context →	Final Exam with Answer Key Instructions: Please carefully read the following passages. For each passage, you must identify which noun the pronoun marked in *bold* refers to. ===== Passage: Mr. Moncrieff visited Chester's luxurious New York apartment, thinking that it belonged to his son Edward. The result was that Mr. Moncrieff has decided to cancel Edward's allowance on the ground that he no longer requires *his* financial support. Question: In the passage above, what does the pronoun " *his* " refer to? Answer:
Target Completion →	mr. moncrieff

Figure G.33: Formatted dataset example for WSC

Context →	Q: 'Nude Descending A Staircase' is perhaps the most famous painting by which 20th century artist? A:
Target Completion →	MARCEL DUCHAMP
Target Completion →	r mutt
Target Completion →	duchamp
Target Completion →	marcel duchamp
Target Completion →	R.Mutt
Target Completion →	Marcel duChamp
Target Completion →	Henri-Robert-Marcel Duchamp
Target Completion →	Marcel du Champ
Target Completion →	henri robert marcel duchamp
Target Completion →	Duchampian
Target Completion →	Duchamp
Target Completion →	duchampian
Target Completion →	marcel du champ
Target Completion →	Marcel Duchamp
Target Completion →	MARCEL DUCHAMP

Figure G.34: Formatted dataset example for TriviaQA. TriviaQA allows for multiple valid completions.

Context →	这赌注让他赢得了四人的晚餐，是关于存在与否的 MASS of the top quark, an elementary particle discovered in 1995. 问题：顶夸克是六种被预言的夸克味中的最后一种 粒子物理的标准模型理论。对还是错？ 答案：
目标补全 →	错误

图G.31: 格式化数据集示例用于RTE

上下文 →	专业装备店提供了一切所需的野生动物园旅行装备。 在他第一次参加步行假期之前，他去了专业装备店购买 一些靴子。 问题：'专业装备店'这个词在上述两个句子中用法是否相同？ 以上句子？ 答案：
目标补全 →	no

图G.32: WiC的格式化数据集示例

上下文 →	期末考试与答案键 说明：请仔细阅读以下段落。 段落，你必须识别加粗的代词指代哪个名词 to。 ===== 段落：Mr. 蒙克雷夫参观了切斯特奢华的纽约公寓， 认为它属于他的儿子爱德华。 蒙克里夫决定取消爱德华的津贴，理由是 他不再需要*他的*金融支持。 问题：在上述段落中，代词"*他的*"指的是什么？ 答案：
目标补全 →	mr. moncrieff

图G.33: 为WSC准备的格式化数据集示例

上下文 →	Q: '裸女下降一座楼梯'或许是哪位 20世纪艺术家的最著名画作？ A:
目标补全 →	马塞尔·杜尚
目标补全 →	r mutt
目标补全 →	杜尚
目标补全 →	马塞尔·杜尚
目标补全 →	R·穆特
目标补全 →	马塞尔·杜尚
目标补全 →	亨利-罗伯特-马塞尔·杜尚
目标补全 →	马塞尔·杜尚
目标补全 →	henri robert marcel duchamp
目标补全 →	杜尚式
目标补全 →	杜尚
目标补全 →	杜尚式
目标补全 →	马塞尔·杜尚
目标补全 →	马塞尔·杜尚
目标补全 →	MARCEL DUCHAMP

图G.34: TriviaQA的格式化数据集示例。TriviaQA允许多个有效补全。

Context	→	Q: What school did burne hogarth establish?
		A:
Target Completion	→	School of Visual Arts

Figure G.35: Formatted dataset example for WebQA

Context	→	Keinesfalls dürfen diese für den kommerziellen Gebrauch verwendet werden. =
Target Completion	→	In no case may they be used for commercial purposes.

Figure G.36: Formatted dataset example for De→En. This is the format for one- and few-shot learning, for this and other language tasks, the format for zero-shot learning is “Q: What is the {language} translation of {sentence} A: {translation}.”

Context	→	In no case may they be used for commercial purposes. =
Target Completion	→	Keinesfalls dürfen diese für den kommerziellen Gebrauch verwendet werden.

Figure G.37: Formatted dataset example for En→De

Context	→	Analysis of instar distributions of larval I. verticalis collected from a series of ponds also indicated that males were in more advanced instars than females. =
Target Completion	→	L’analyse de la distribution de fréquence des stades larvaires d’I. verticalis dans une série d’étangs a également démontré que les larves mâles étaient à des stades plus avancés que les larves femelles.

Figure G.38: Formatted dataset example for En→Fr

Context	→	L’analyse de la distribution de fréquence des stades larvaires d’I. verticalis dans une série d’étangs a également démontré que les larves mâles étaient à des stades plus avancés que les larves femelles. =
Target Completion	→	Analysis of instar distributions of larval I. verticalis collected from a series of ponds also indicated that males were in more advanced instars than females.

Figure G.39: Formatted dataset example for Fr→En

Context	→	The truth is that you want, at any price, and against the wishes of the peoples of Europe, to continue the negotiations for Turkey’s accession to the European Union, despite Turkey’s continuing refusal to recognise Cyprus and despite the fact that the democratic reforms are at a standstill. =
Target Completion	→	Adevărul este că vă doriți, cu orice preț și împotriva dorinței europenilor, să continuați negocierile de aderare a Turciei la Uniunea Europeană, în ciuda refuzului continuu al Turciei de a recunoaște Ciprul și în ciuda faptului că reformele democratice au ajuns într-un punct mort.

Figure G.40: Formatted dataset example for En→Ro

Context	→	Q: 班恩·霍格思在哪个学校创立?
		A:
Target 补全	→	视觉艺术学校

图G.35: 格式化数据集示例用于WebQA

Context	→	绝对禁止将此f用于商业用途。 =
Target 补全	→	在任何情况下都不得将其用于商业目的。

图G.36: 格式化数据集示例用于De→En。这是用于一元和少样本学习 (few-shot learning) 的格式，对于这个和其他语言任务，零样本学习 (zero-shot learning) 的格式是“Q: {语言} 翻译 {句子} A: {翻译}。”

Context	→	在任何情况下都不得将其用于商业用途。 =
目标补全	→	绝对不能将这些f用于商业用途。

图G.37: 为En→De准备的格式化数据集示例

上下文	→	对从一系列池塘采集的垂直鳞翅目幼虫I. verticalis的分析也表明雄性处于更高级的龄期比女性多。 =
目标补全	→	对I. verticalis幼虫期频率分布的分析垂直is在一系列 etangs中同样d emontr e表明, 雄性larves mâles etaient a处于比雌性larves更avanc es的stades.

图G.38: 格式化数据集示例用于En→Fr

上下文	→	I. verticalis 幼虫阶段频率分布分析在一系列 池中, 也证明 了幼 雌性幼虫处于更先进的阶段雄性幼虫 已发育到比雌性幼虫更 的阶段 =
目标补全	→	从 I. verticalis 幼虫收集的幼期分布分析一系列池塘也表明, 雄性处于更高级的龄期, 而雌性则不是。

图G.39: Fr→En的格式化数据集示例

上下文	→	事实是, 无论付出什么代价, 都要违背欧洲各国的意愿, 继续土耳其加入的谈判, 对欧盟, 尽管土耳其持续拒绝承认塞浦路斯, 并且民主改革正处于停滞状态停滞不前。 =
目标补全	→	Adev arul este că a v a doriți, cu orice pret. si împotriva dorinței europenilor, să continuați negocierile de aderare a Turciei la Uniunea Europeană, 尽管土耳其持续拒绝承认塞浦路斯和 尽管民主改革已陷入僵局

图G.40: 为En→Ro准备的格式化数据集示例

Context →	Adevărul este că vă doriți, cu orice preț și împotriva dorinței europenilor, să continuați negocierile de aderare a Turciei la Uniunea Europeană, în ciuda refuzului continuu al Turciei de a recunoaște Ciprul și în ciuda faptului că reformele democratice au ajuns într-un punct mort.
-----------	--

Target Completion →	The truth is that you want, at any price, and against the wishes of the peoples of Europe, to continue the negotiations for Turkey's accession to the European Union, despite Turkey's continuing refusal to recognise Cyprus and despite the fact that the democratic reforms are at a standstill.
---------------------	---

Figure G.41: Formatted dataset example for Ro→En

Context →	Q: What is (2 * 4) * 6? A:
-----------	-------------------------------

Target Completion →	48
---------------------	----

Figure G.42: Formatted dataset example for Arithmetic 1DC

Context →	Q: What is 17 minus 14? A:
-----------	-------------------------------

Target Completion →	3
---------------------	---

Figure G.43: Formatted dataset example for Arithmetic 2D-

Context →	Q: What is 98 plus 45? A:
-----------	------------------------------

Target Completion →	143
---------------------	-----

Figure G.44: Formatted dataset example for Arithmetic 2D+

Context →	Q: What is 95 times 45? A:
-----------	-------------------------------

Target Completion →	4275
---------------------	------

Figure G.45: Formatted dataset example for Arithmetic 2Dx

Context →	Q: What is 509 minus 488? A:
-----------	---------------------------------

Target Completion →	21
---------------------	----

Figure G.46: Formatted dataset example for Arithmetic 3D-

Context →	Q: What is 556 plus 497? A:
-----------	--------------------------------

Target Completion →	1053
---------------------	------

Figure G.47: Formatted dataset example for Arithmetic 3D+

Context →	Q: What is 6209 minus 3365? A:
-----------	-----------------------------------

Target Completion →	2844
---------------------	------

Figure G.48: Formatted dataset example for Arithmetic 4D-

上下文 →	成功就是想要什么, 不惜任何代价反对想要的东西 对欧洲人来说, 继续进行土耳其加入欧盟的谈判 欧洲, 尽管土耳其持续拒绝承认塞浦路斯 而且, 尽管民主改革已经陷入僵局
-------	--

目标补全 →	事实是, 你愿意不惜一切代价, 不顾欧洲人民的意愿, 继续就土耳其加入欧盟的谈判 欧洲人民, 继续就土耳其加入进行谈判 尽管土耳其持续拒绝承认塞浦路斯, 尽管民主改革陷入停滞 尽管民主改革陷入停滞 停滞。
--------	--

图G.41: 格式化数据集示例 →En

上下文 →	Q: 什么是 (2 * 4) * 6? A:
-------	---------------------------

目标补全 →	48
--------	----

图 G.42: 用于算术 1DC 的格式化数据集示例

上下文 →	Q: 什么是 17 减 14? A:
-------	-----------------------

目标补全 →	3
--------	---

图G.43: 用于算术2D-的格式化数据集示例

上下文 →	Q: 98加45等于多少? A:
-------	---------------------

目标补全 →	143
--------	-----

图G.44: 用于算术2D+的格式化数据集示例

上下文 →	Q: 95 乘以 45 等于多少? A:
-------	-------------------------

目标补全 →	4275
--------	------

图 G.45: 算术2Dx的格式化数据集示例

上下文 →	Q: 509 减去 488 等于多少? A:
-------	---------------------------

目标补全 →	21
--------	----

图G.46: 算术3D-格式化数据集示例

上下文 →	Q: 556加497等于多少? A:
-------	-----------------------

目标补全 →	1053
--------	------

图G.47: 算术3D+格式化数据集示例

上下文 →	Q: 6209 减去 3365 是多少? A:
-------	----------------------------

目标补全 →	2844
--------	------

图 G.48: 算术 4D- 的格式化数据集示例

Context →	Q: What is 9923 plus 617? A:
Target Completion →	10540

Figure G.49: Formatted dataset example for Arithmetic 4D+

上下文 →	Q: 9923 加 617 等于多少? A:
目标补全 →	10540

图 G.49: 算术 4D 的格式化数据集示例+

Context →	Q: What is 40649 minus 78746? A:
Target Completion →	-38097

Figure G.50: Formatted dataset example for Arithmetic 5D-

上下文 →	Q: 40649 减去 78746 等于多少? A:
目标补全 →	-38097

图G.50: 用于算术5D的格式化数据集示例-

Context →	Q: What is 65360 plus 16204? A:
Target Completion →	81564

Figure G.51: Formatted dataset example for Arithmetic 5D+

上下文 →	Q: 65360 加 16204 等于多少? A:
目标补全 →	81564

图G.51: 用于算术5D的格式化数据集示例+

H Results on All Tasks for All Model Sizes

Name	Metric	Split	Fine-tune SOTA K	Zero-Shot								One-Shot								Few-Shot								175B (test server)		
				Small	Med	Large	XL	2.7B	6.7B	13B	175B	Small	Med	Large	XL	2.7B	6.7B	13B	175B	Small	Med	Large	XL	2.7B	6.7B	13B	175B			
HellaSwag	acc	dev	85.6	20	33.7	43.6	51.0	54.7	62.8	67.4	70.9	78.9	33.0	42.9	50.5	53.5	61.9	66.5	70.0	78.1	33.5	43.1	51.3	54.9	62.9	67.3	71.3	79.3		
LAMBADA	acc	test	68.0	15	42.7	54.3	60.4	63.6	67.1	70.3	72.5	76.2	22.0	47.1	52.6	58.3	61.1	65.4	69.0	72.5	22.0	40.4	63.2	57.0	78.1	79.1	81.3	86.4		
LAMBADA	ppl	test	8.63	15	18.6	9.09	6.53	5.44	4.60	4.00	3.56	3.00	165.0	11.6	8.29	6.46	5.53	4.61	4.06	3.35	165.0	27.6	6.63	7.45	2.89	2.56	2.56	1.92		
StoryCloze	acc	test	91.8	70	63.3	68.5	72.4	73.4	77.2	77.7	79.5	83.2	62.3	68.7	72.3	74.2	77.3	78.7	79.7	84.7	62.3	70.2	73.9	76.1	80.2	81.2	83.0	87.7		
NQs	acc	test	44.5	64	0.64	1.75	2.71	4.40	6.01	5.79	7.84	14.6	1.19	3.07	4.79	5.43	8.73	9.78	13.7	23.0	1.72	4.46	7.89	9.72	13.2	17.0	21.0	29.9		
TriviaQA	acc	dev	68.0	64	4.15	7.61	14.0	19.7	31.3	38.7	41.8	64.3	4.19	12.9	20.5	26.5	35.9	44.4	51.3	68.0	6.96	16.3	26.5	32.1	42.3	51.6	57.5	71.2	71.2	
Webqs	acc	test	45.5	64	1.77	3.20	4.33	4.63	7.92	7.73	8.22	14.4	2.56	6.20	8.51	9.15	14.5	15.1	19.0	25.3	5.46	12.6	15.9	19.6	24.8	27.7	33.5	41.5		
Ro→En 16	BLEU-mb	test	39.9	64	2.08	2.71	3.09	3.15	16.3	8.34	20.2	19.9	0.55	15.4	23.0	26.3	30.6	33.2	35.6	38.6	1.25	20.7	25.8	29.2	33.1	34.8	37.0	39.5		
Ro→En 16	BLEU-sb	test	64	64	2.39	3.08	3.49	3.56	16.8	8.75	20.8	20.9	0.65	15.9	23.6	26.8	31.3	34.2	36.7	40.0	1.40	21.3	26.6	30.1	34.3	36.2	38.4	41.3		
En→Ro 16	BLEU-mb	test	38.5	64	2.14	2.65	2.53	2.50	3.46	4.24	5.32	14.1	0.35	3.30	7.89	8.72	13.2	15.1	17.3	20.6	1.25	5.90	9.33	10.7	14.3	16.3	18.0	21.0		
En→Ro 16	BLEU-sb	test	64	64	2.61	3.11	3.07	3.09	4.26	5.31	6.43	18.0	0.55	3.90	9.15	10.3	15.7	18.2	20.8	24.9	1.64	7.40	10.9	12.9	17.2	19.6	21.8	25.8		
Fr→En 14	BLEU-mb	test	35.0	64	1.81	2.53	3.47	3.13	20.6	15.1	21.8	21.2	1.28	15.9	23.7	26.3	29.0	30.5	30.2	33.7	4.98	25.5	28.5	31.1	33.7	34.9	36.6	39.2		
Fr→En 14	BLEU-sb	test	64	64	2.29	2.99	3.90	3.60	21.2	15.5	22.4	21.9	1.50	16.3	24.4	27.0	30.0	31.6	31.4	35.6	5.30	26.2	29.5	32.2	35.1	36.4	38.3	41.4		
En→Fr 14	BLEU-mb	test	45.6	64	1.74	2.16	2.73	2.15	15.1	8.82	12.0	25.2	0.49	8.00	14.8	15.9	20.3	23.3	24.9	28.3	4.08	14.5	19.3	21.5	24.9	27.3	29.5	32.6		
En→Fr 14	BLEU-sb	test	45.9	64	2.44	2.75	3.54	2.82	19.3	11.4	15.3	31.3	0.81	10.0	18.2	19.3	24.7	28.3	30.1	34.1	5.31	18.0	23.6	26.1	30.3	33.3	35.5	39.9		
De→En 16	BLEU-mb	test	40.2	64	2.06	2.87	3.41	3.63	21.5	17.3	23.0	27.2	0.83	16.2	22.5	24.7	28.2	30.7	33.0	30.4	3.25	22.7	26.2	29.2	32.7	34.8	37.3	40.6		
De→En 16	BLEU-sb	test	64	64	2.39	3.27	3.85	4.04	22.5	18.2	24.4	28.6	0.93	17.1	23.4	25.8	29.2	31.9	34.5	32.1	3.60	23.8	27.5	30.5	34.1	36.5	39.3	43.0		
En→De 16	BLEU-mb	test	41.2	64	1.70	2.27	2.31	2.43	12.9	8.66	10.4	24.6	0.50	7.00	12.9	13.1	18.3	20.9	22.5	26.2	3.42	12.3	15.4	17.1	20.9	23.0	26.6	29.7		
En→De 16	BLEU-sb	test	41.2	64	2.09	2.65	2.75	2.92	13.7	9.36	11.0	25.3	0.54	7.40	13.4	13.4	18.8	21.7	23.3	27.3	3.78	12.9	16.1	17.7	21.7	24.1	27.7	30.9		
Winograd	acc	test	93.8	7	66.3	72.9	74.7	76.9	82.4	85.7	87.9	88.3	63.4	68.5	72.9	76.9	82.4	84.6	86.1	89.7	63.4	67.4	73.6	76.9	84.3	85.4	82.4	88.6		
Winogrande	acc	dev	84.6	50	52.1	57.4	58.7	62.3	64.5	67.0	70.0	77.2	51.3	52.6	57.5	59.1	61.7	65.8	66.9	73.2	51.3	52.6	57.5	59.1	62.6	67.4	70.0	77.2		
PIQA	acc	dev	77.1	50	64.6	70.2	72.9	75.1	75.6	78.0	78.5	81.0	64.3	69.3	71.8	74.4	74.3	76.3	77.8	80.5	64.3	69.4	72.0	74.3	75.4	77.8	79.9	82.3		
ARC (Challenge)	acc	test	78.5	50	26.6	29.5	31.8	35.5	38.0	41.4	43.7	51.4	25.5	30.2	31.6	36.4	38.4	41.5	43.1	53.2	25.5	28.4	32.3	36.7	39.5	43.7	44.8	51.5		
ARC (Easy)	acc	test	92.0	50	43.6	46.5	53.0	53.8	58.2	60.2	63.8	68.8	42.7	48.2	54.6	55.9	60.3	62.6	66.8	71.2	42.7	48.2	54.6	55.9	60.3	62.6	66.8	71.2		
OpenBookQA	acc	test	87.2	100	35.6	43.2	45.2	46.8	53.0	50.4	55.6	57.6	37.0	39.8	46.2	46.4	53.4	53.0	55.8	58.8	37.0	43.6	48.0	46.4	53.4	53.0	55.8	58.8		
Quac	f1	dev	74.4	5	21.2	26.8	31.0	30.1	34.7	36.1	38.4	41.5	21.1	26.9	31.9	32.3	37.4	39.0	40.6	43.4	21.6	27.6	32.9	34.2	38.2	39.9	40.9	44.3		
RACE-h	acc	test	90.0	10	35.2	37.9	40.1	40.9	42.4	44.1	44.6	45.5	34.3	37.7	40.0	42.0	43.8	44.3	44.6	45.9	34.3	37.0	40.4	41.4	42.3	44.7	45.1	46.8		
RACE-m	acc	test	93.1	10	42.1	47.2	52.1	52.3	54.7	54.4	56.7	58.4	42.3	47.3	51.7	55.2	56.1	54.7	56.9	57.4	42.3	47.0	52.7	52.3	54.7	54.4	56.7	58.4		
SQuADv2	em	dev	90.7	16	22.6	32.8	33.9	43.1	43.6	45.4	49.0	52.6	25.1	37.5	37.9	47.9	47.9	51.1	56.0	60.1	27.5	40.5	39.2	53.5	50.0	56.6	62.6	64.9		
SQuADv2	f1	dev	93.0	16	28.3	40.2	41.4	50.3	51.0	52.7	56.3	59.5	32.1	43.6	44.1	54.0	54.1	57.1	61.8	65.4	32.1	43.6	44.1	54.0	54.1	57.1	61.8	65.4		
CoQA	f1	dev	90.7	5	34.5	55.0	61.8	65.3	71.1	72.8	76.3	81.5	30.6	52.1	61.6	66.1	71.8	75.1	77.9	84.0	31.1	52.0	62.7	66.8	73.2	77.3	79.9	85.0		
DROP	f1	dev	89.1	20	9.40	13.6	14.4	16.4	19.7	17.0	20.4	23.6	11.7	18.1	20.9	23.0	26.4	27.3	29.2	34.3	12.9	18.7	24.0	25.6	29.7	29.7	32.3	36.5		
BoolQ	acc	dev	91.0	32	49.7	60.3	58.9	62.4	67.1	65.4	66.2	60.5	52.6	61.7	60.4	63.7	68.4	68.7	69.0	76.7	43.1	60.6	62.0	62.4	67.1	65.4	66.2	60.5		
CB	acc	dev	96.9	32	0.00	32.1	8.93	19.6	19.6	28.6	19.6	46.4	55.4	53.6	53.6	48.2	57.1	33.9	55.4	64.3	42.9	58.9	53.6	69.6	67.9	60.7	66.1	82.1		
CB	f1	dev	93.9	32	0.00	29.3	11.4	17.4	22.4	25.1	20.3	42.8	60.1	40.4	32.6	48.3	45.7	44.6	46.0	57.2	52.0	61.0	32.6	48.3	45.7	44.6	46.0	57.2		
Copa	acc	dev	94.8	32	66.0	68.0	73.0	77.0	76.0	80.0	84.0	91.0	62.0	64.0	66.0	77.0	83.0	83.0	86.0	92.0	92.0	66.0	68.0	73.0	77.0	76.0	80.0	84.0	91.0	
RTE	acc	dev	92.5	32	47.7	49.8	48.4	56.0	46.6	55.2	62.8	63.5	53.1	47.3	49.5	49.5	54.9	54.9	56.3	70.4	52.3	48.4	46.9	50.9	56.3	49.5	60.6	72.9		
WiC	acc	dev	76.1	32	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	50.0	0.00	0.00	53.0	49.2	49.4	50.3	50.0	49.4	49.4	50.0	0.00	0.00	0.00	0.00	0.00		
WSC	acc	dev	93.8	32	59.6	56.7	65.4	61.5	66.3	60.6	64.4	65.4	58.7	58.7	60.6	62.5	66.3	60.6	66.3	69.2	58.7	60.6	54.8	62.5	66.3	60.6	66.3	69.2		
MultiRC	acc	dev	62.3	32	4.72	9.65	12.3	13.6	14.3	18.4	24.2	27.6	4.72	9.65	12.3	13.6	14.3	18.4	24.2	27.6	6.09	11.8	16.8	20.8	24.7	23.8	25.0	32.5		
MultiRC	f1a	dev	88.2	32	57.0	59.7	60.4	59.9	60.0	64.5	71.4	72.9	57.0	59.7	60.4	59.9	60.0	64.5	71.4	72.9										

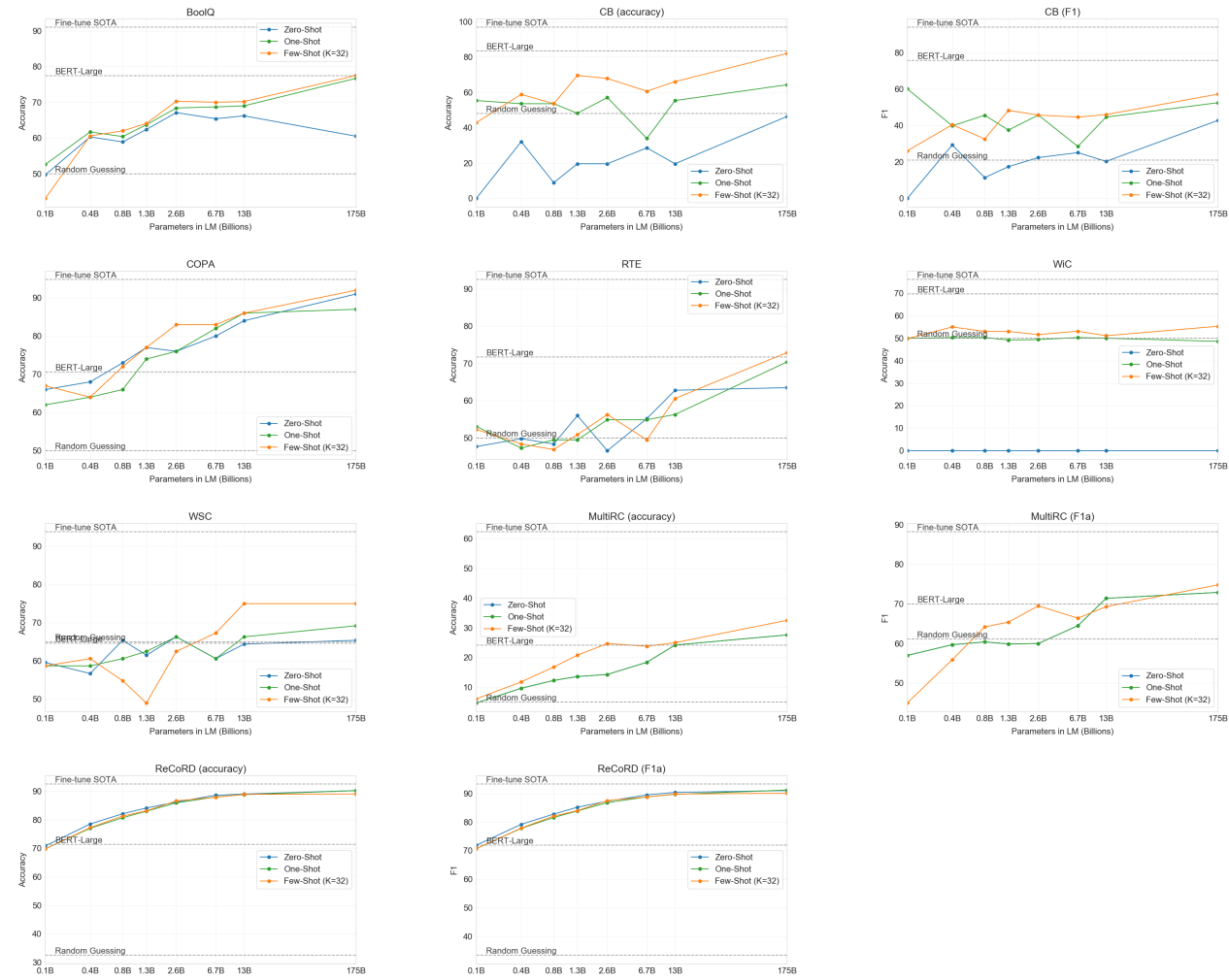


Figure H.1: All results for all SuperGLUE tasks.

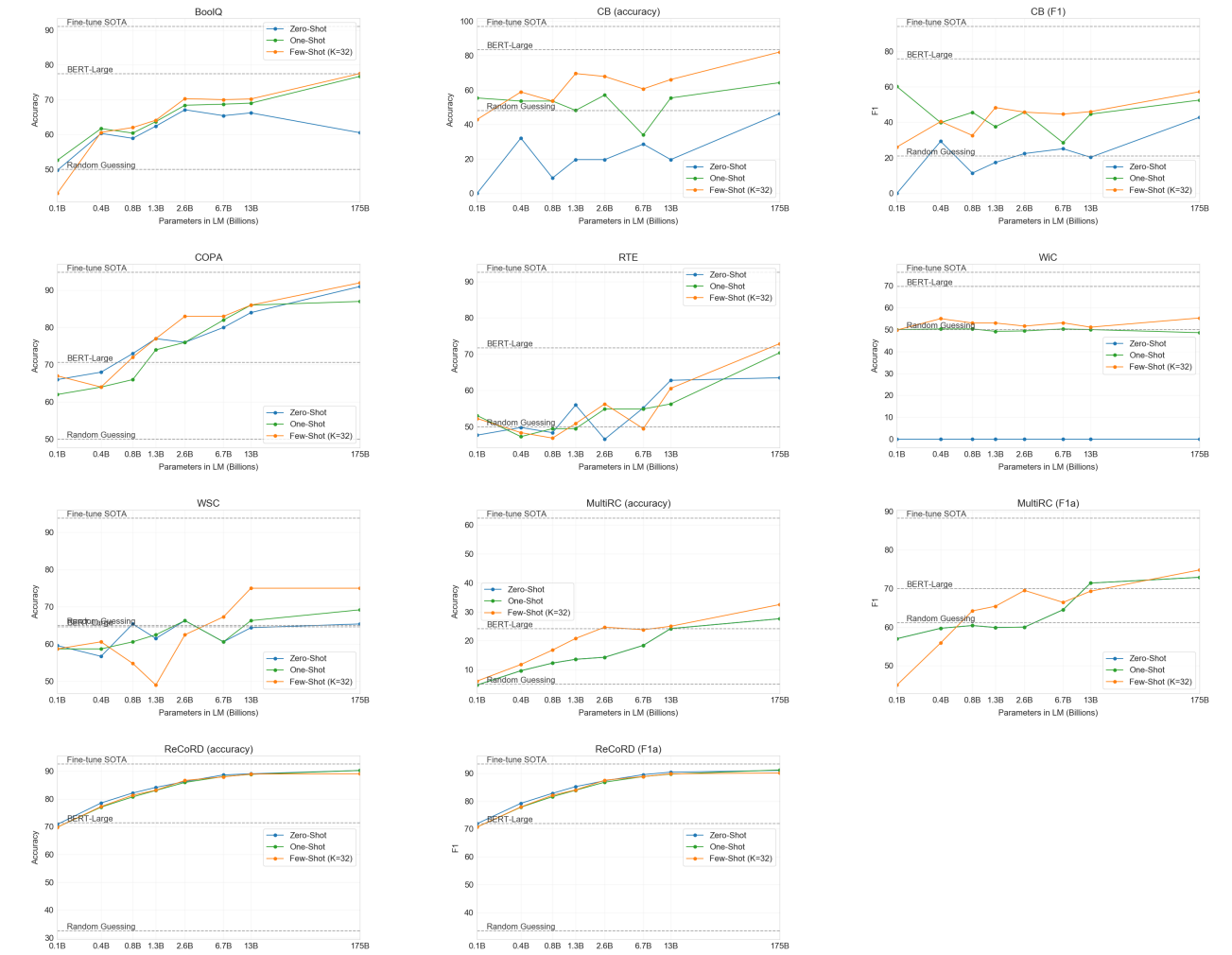


图 H.1: 所有 SuperGLUE 任务的测试结果。

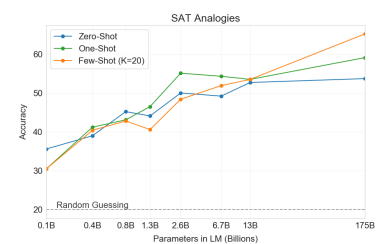


Figure H.2: Results for SAT task.

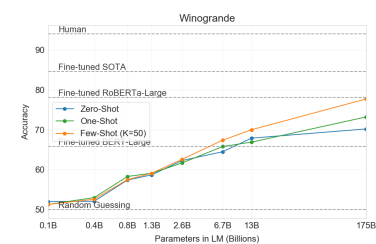


Figure H.3: All results for all Winogrande tasks.

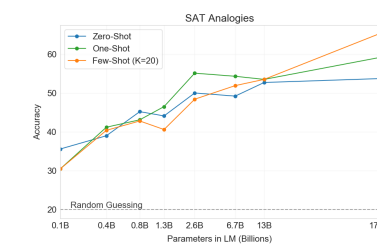
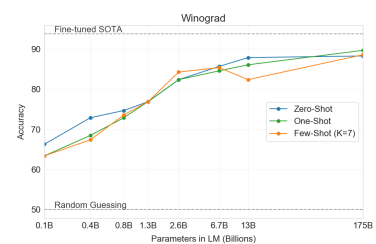


图 H.2: SAT 任务的测试结果。



图 H.3: 所有 Winograd 任务的测试结果。

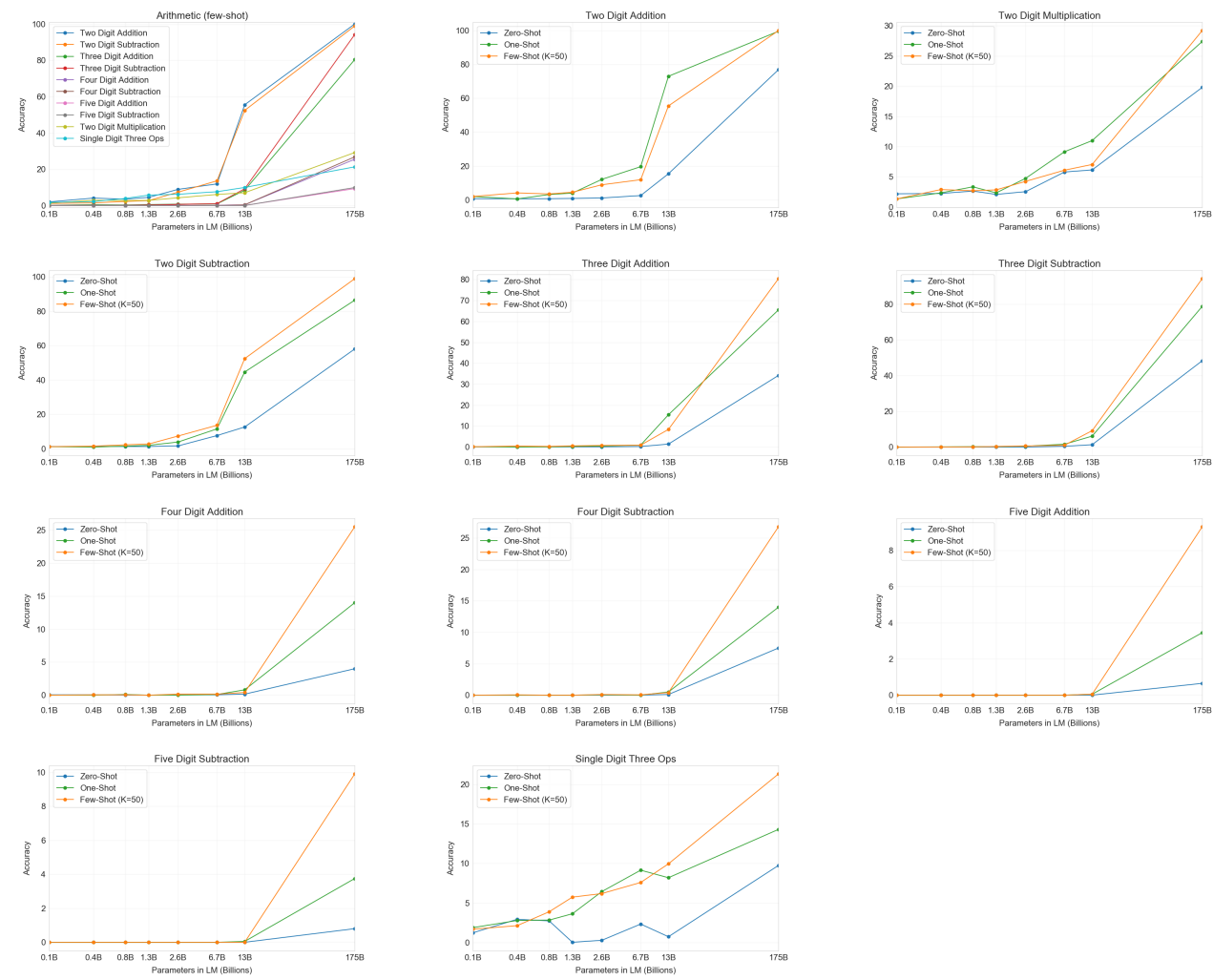


Figure H.4: All results for all Arithmetic tasks.

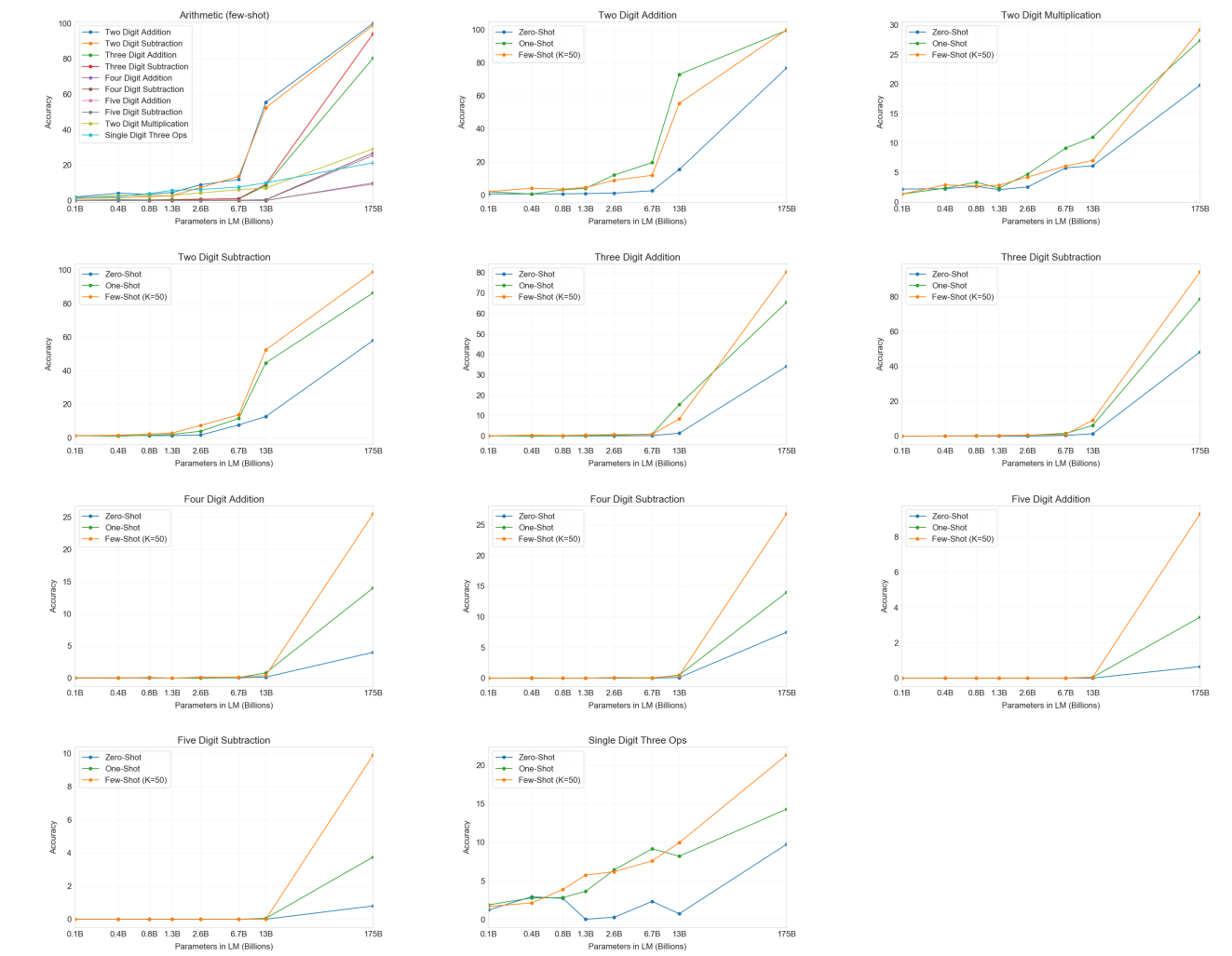


图 H.4: 所有算术任务的结果。

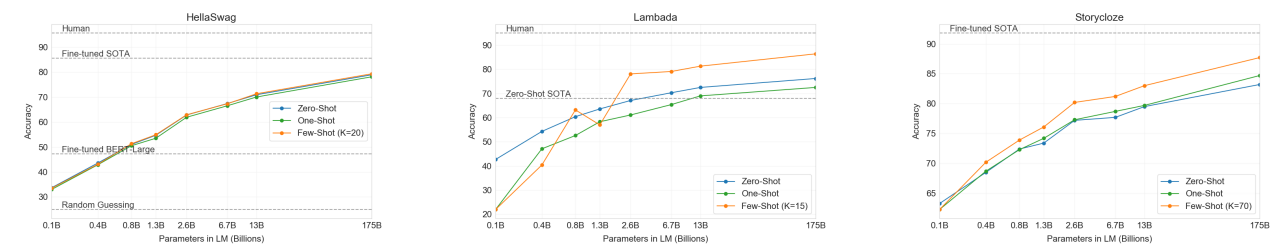


Figure H.5: All results for all Cloze and Completion tasks.

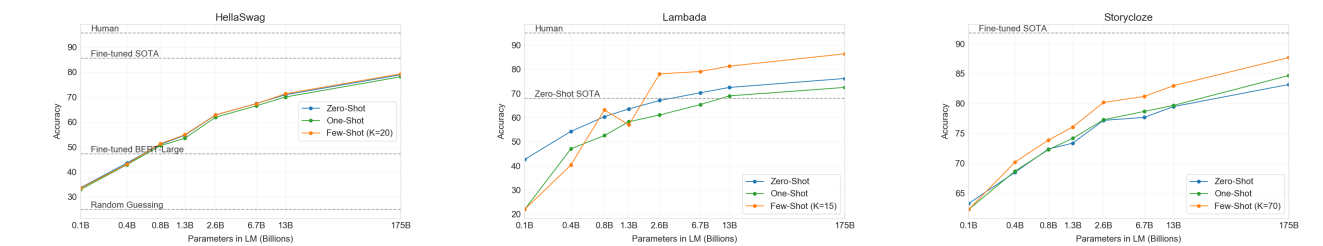


图 H.5: 所有完形填空和补全任务的结果。

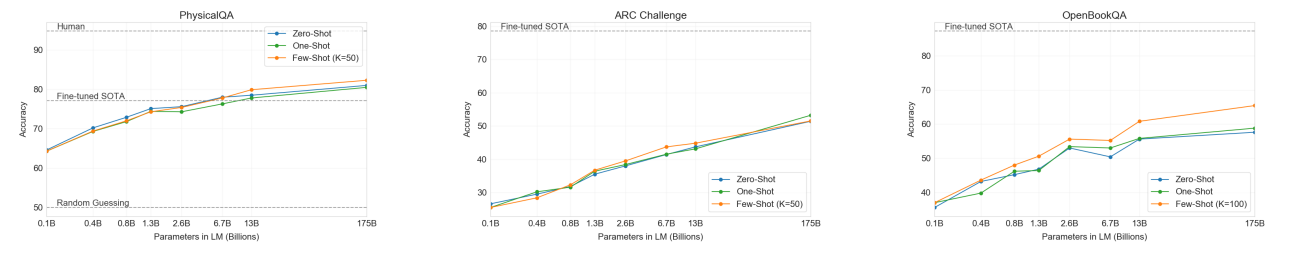
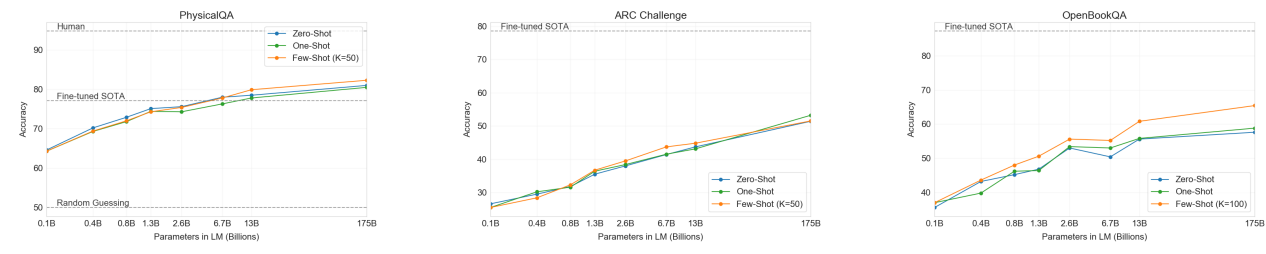


Figure H.6: All results for all Common Sense Reasoning tasks.

图 H.6: 所有常识推理任务的全部结果。

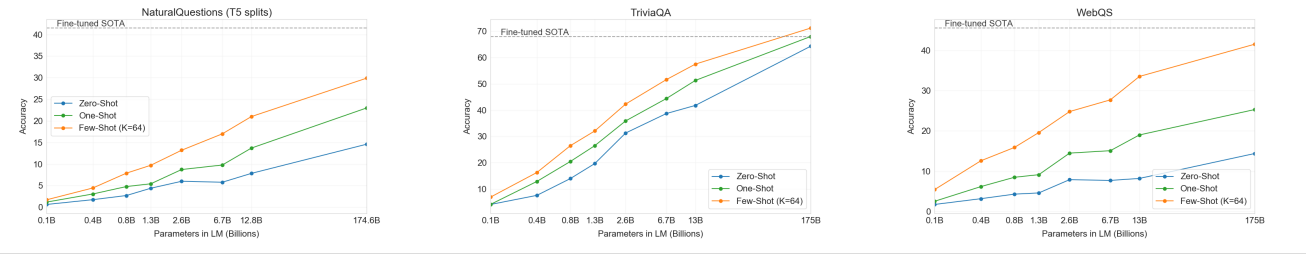
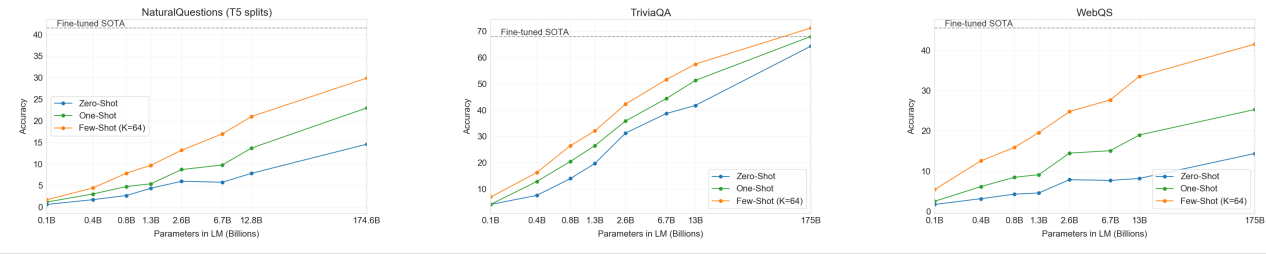


Figure H.7: All results for all QA tasks.

图 H.7: 所有问答任务的全部结果。

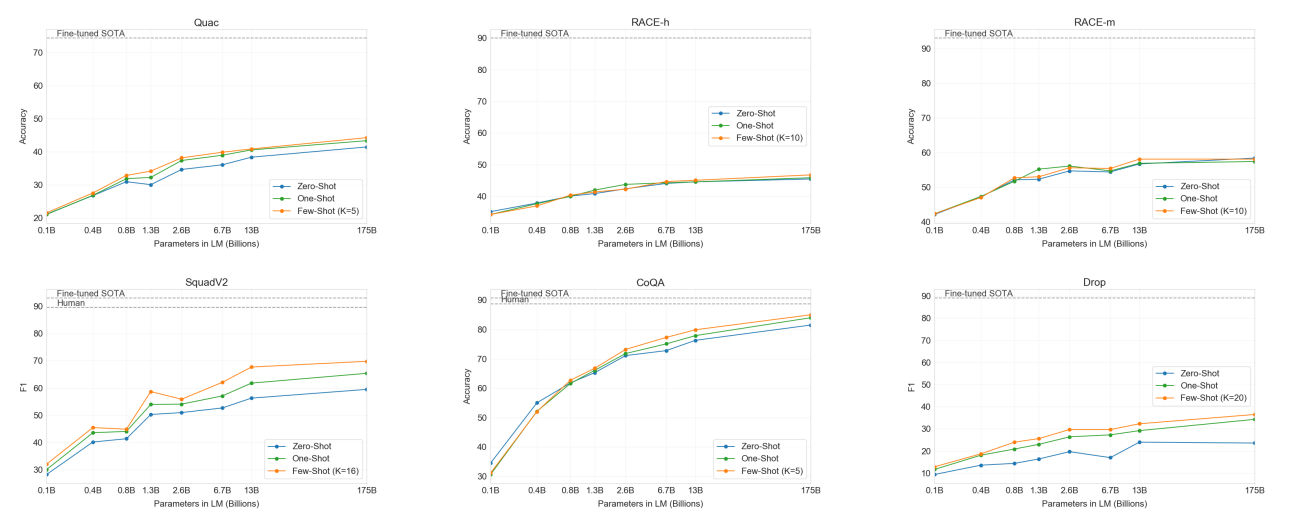
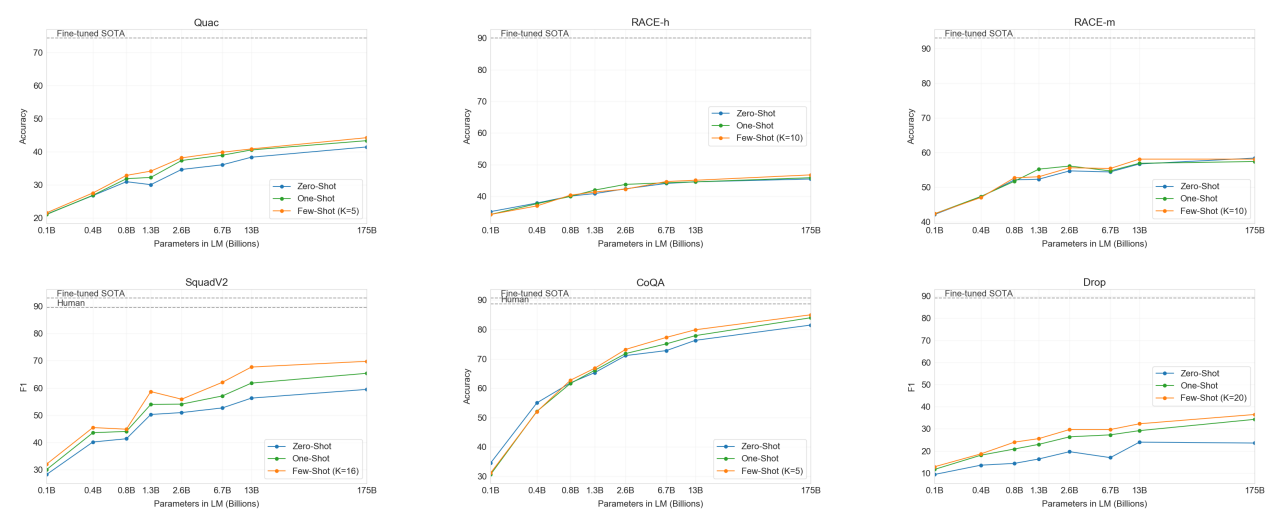


Figure H.8: All results for all Reading Comprehension tasks.

图 H.8: 所有阅读理解任务的全部结果。

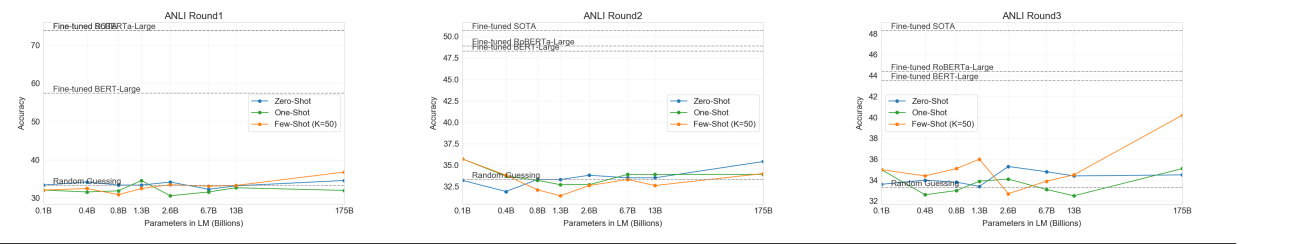
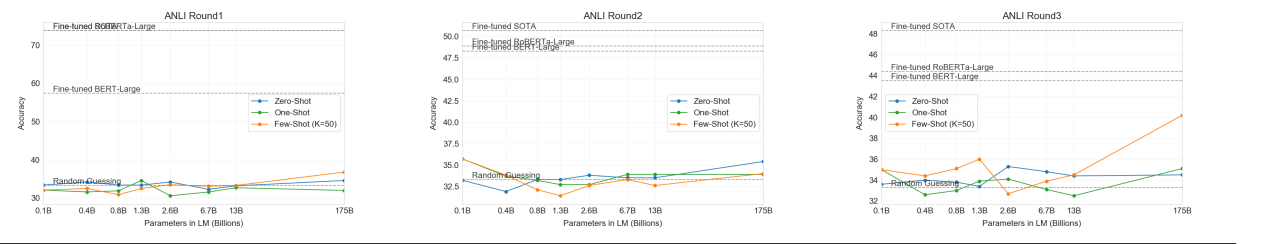


Figure H.9: All results for all ANLI rounds.

图 H.9: 所有 ANLI 轮次的全部结果。

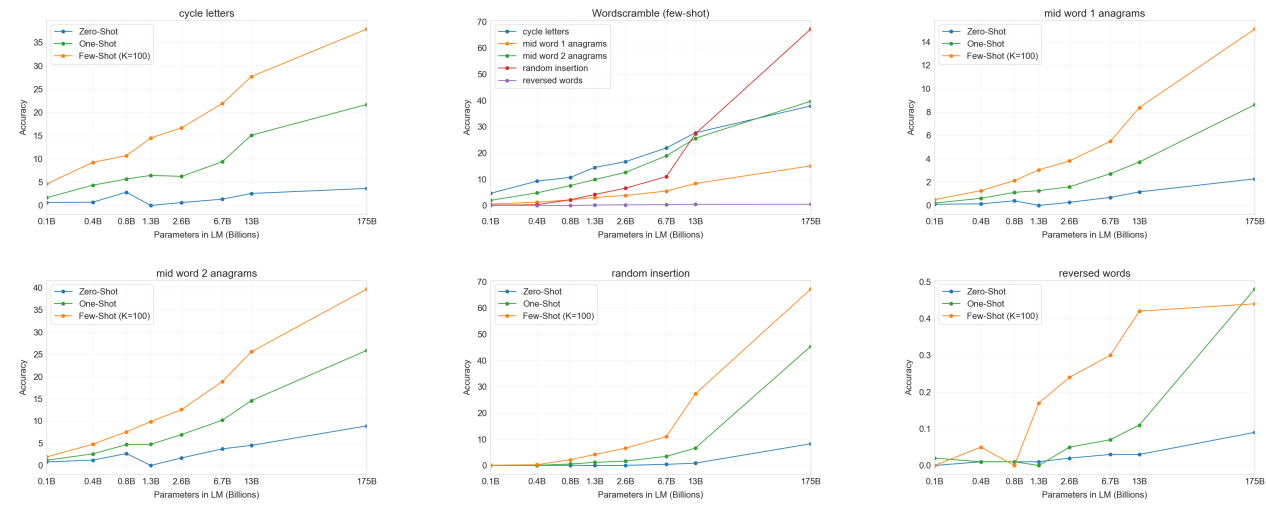


Figure H.10: All results for all Scramble tasks.

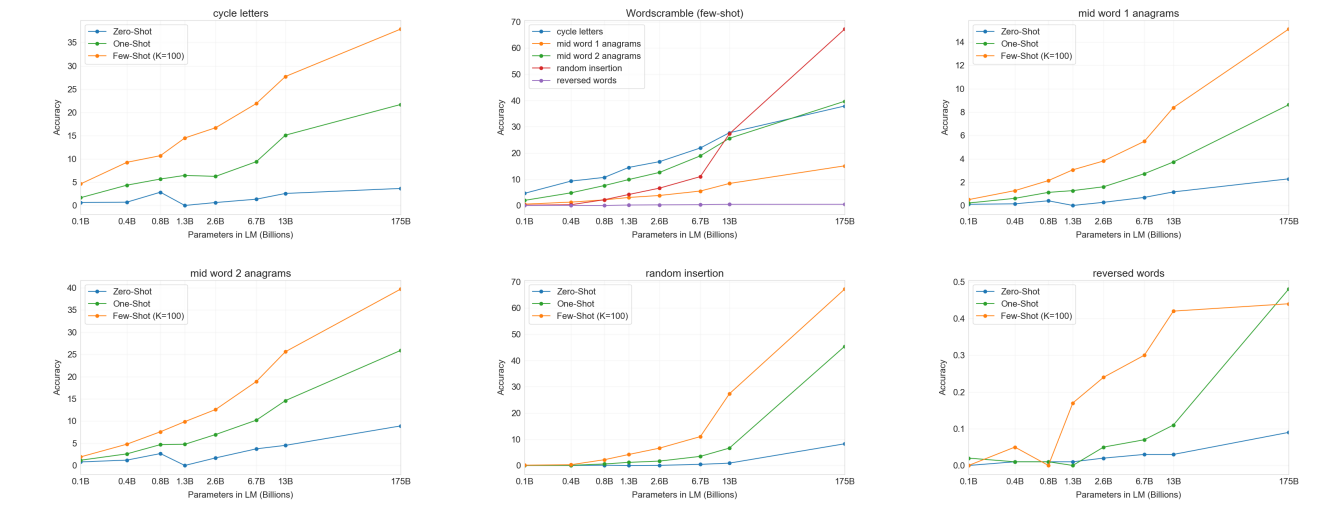


图 H.10: 所有打乱任务的全部结果。

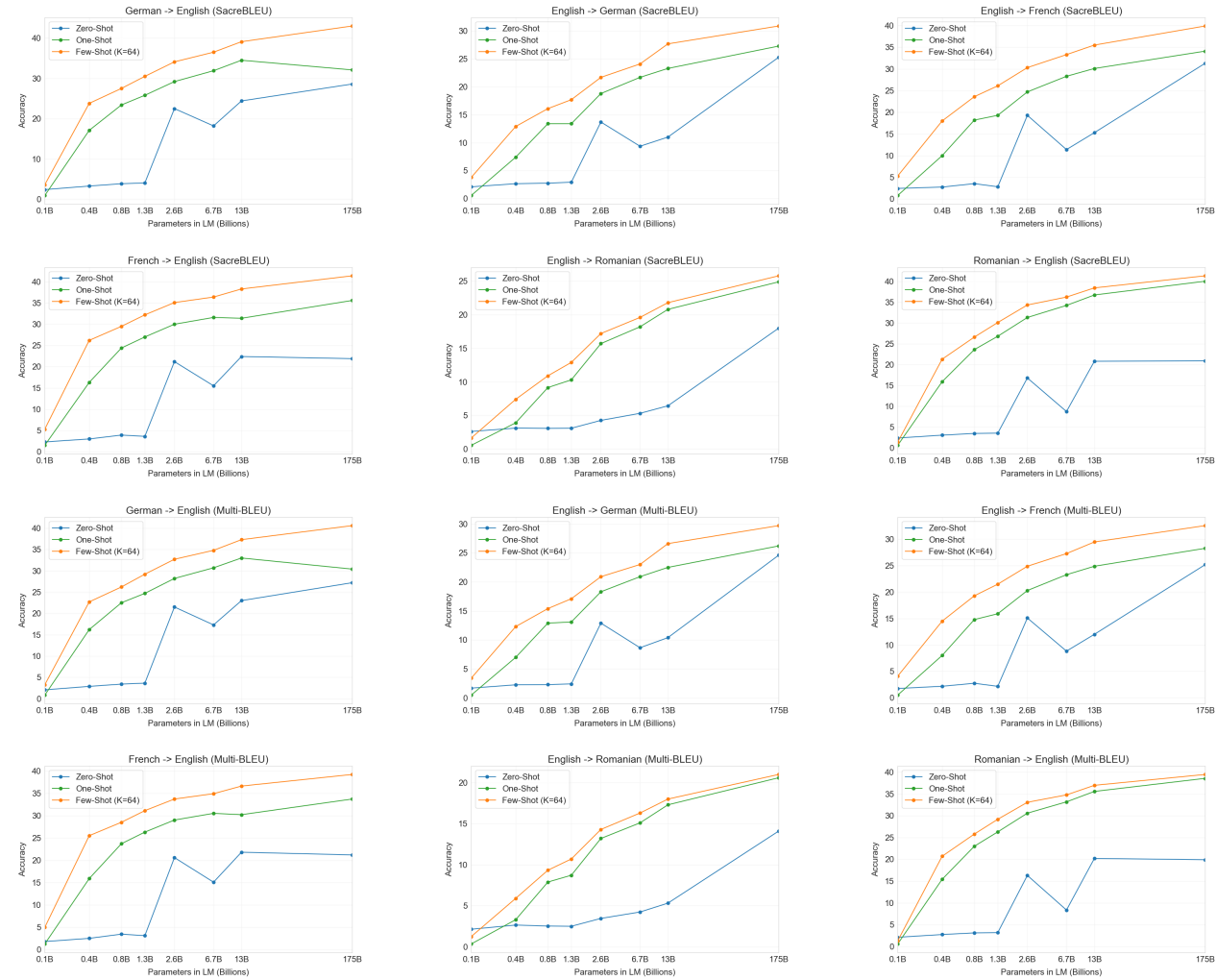


Figure H.11: All results for all Translation tasks.

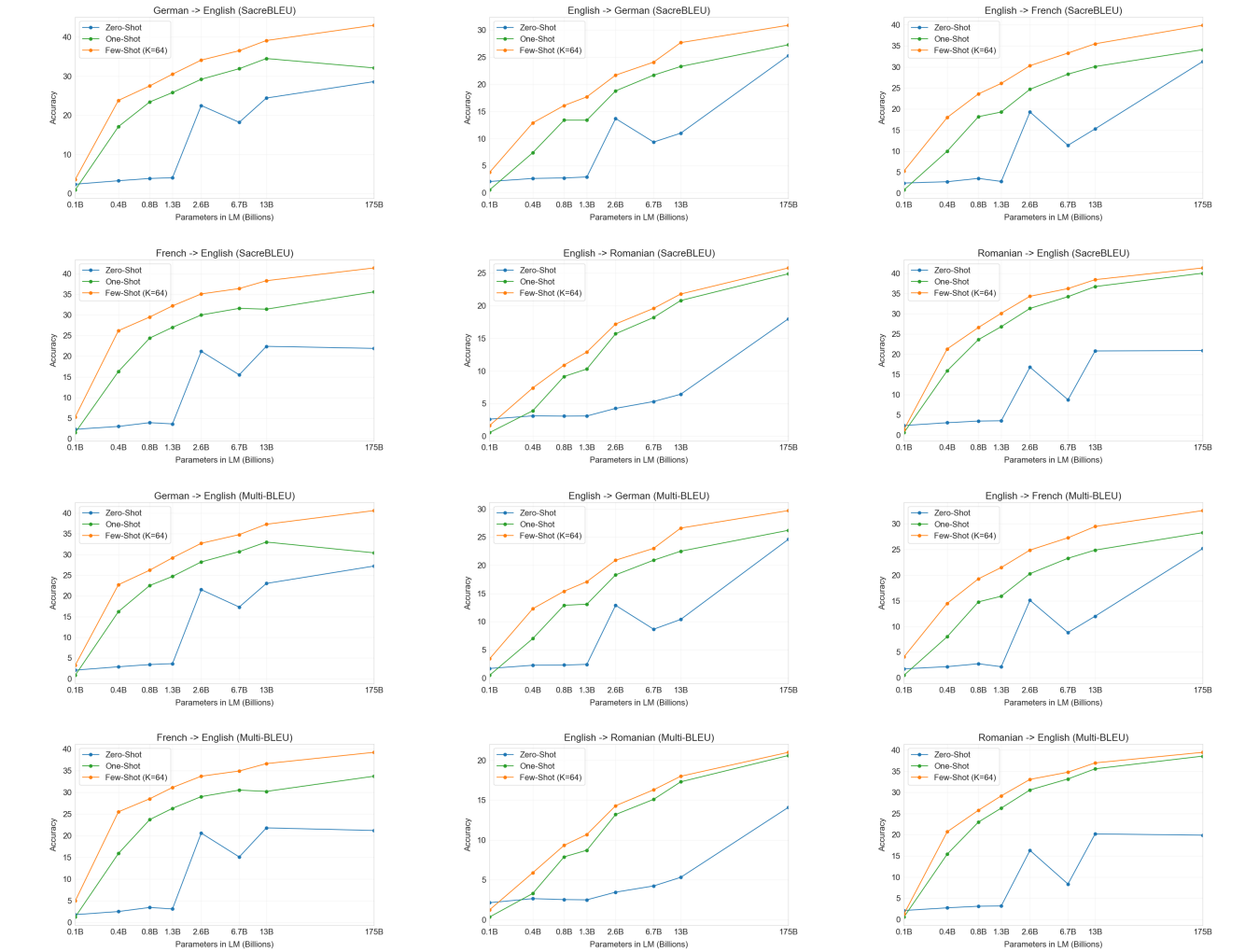


图 H.11: 所有翻译任务的全部结果。

References

- [ADG⁺16] Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando De Freitas. Learning to learn by gradient descent by gradient descent. In *Advances in neural information processing systems*, pages 3981–3989, 2016.
- [AI19] WeChat AI. Tr-mt (ensemble), December 2019.
- [AJF19] Roei 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.
- [BCFL13] Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1533–1544, 2013.
- [BDD⁺09] Luisa Bentivogli, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, and Bernardo Magnini. The fifth PASCAL recognizing textual entailment challenge. 2009.
- [BES10] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, volume 10, pages 2200–2204, 2010.
- [BHDD⁺06] Roy Bar Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. The second PASCAL recognising textual entailment challenge. 2006.
- [BHT⁺20] Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. Experience grounds language. *arXiv preprint arXiv:2004.10151*, 2020.
- [BLC13] Yoshua Bengio, Nicholas Léonard, and Aaron C. Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *Arxiv*, 2013.
- [BZB⁺19] Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. *arXiv preprint arXiv:1911.11641*, 2019.
- [Car97] Rich Caruana. Multitask learning. *Machine learning*, 28(1), 1997.
- [CB78] Susan Carey and Elsa Bartlett. Acquiring a single new word. *Proceedings of the Stanford Child Language Conference*, 1978.
- [CCE⁺18] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457, 2018.
- [CGRS19] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers, 2019.
- [CHI⁺18] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. Quac : Question answering in context. *Arxiv*, 2018.
- [CLC⁺19] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*, 2019.
- [CLY⁺19] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740*, 2019.
- [Cra17] Kate Crawford. The trouble with bias. *NIPS 2017 Keynote*, 2017.
- [DCLT18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

参考文献

- [ADG⁺16] 亚历克·安德鲁希奥维兹, 米哈伊尔·德尼尔, 塞尔吉奥·戈麦斯, 马修·W·霍夫曼, 大卫·普劳, 汤姆·绍尔, 布兰登·希林福德, 以及南多·德·弗雷塔斯。通过梯度下降学习如何学习。在神经信息处理系统进展, 第3981–3989页, 2016年.[AI19] 微信AI. Tr-mt (集成), 2019年12月.[AJF19] 罗伊·阿哈罗尼, 梅尔文·约翰逊, 和奥汉·费拉特。大规模多语言神经机器翻译。在2019年北美计算语言学协会会议论文集: 人机语言技术, 第一卷(长篇和短篇论文), 2019年.[BBDIW20] 苏·林·布洛德盖特, 索伦·巴罗卡斯, 哈尔·道姆 e III, 和汉娜·瓦拉奇。语言(技术)就是力量: 对nlp中“偏见”的批判性调查。arXiv预印本 arXiv:2005.14050, 2020年.[BCFL13] 乔纳森·贝尔兰特, 安德鲁·周, 罗伊·弗罗斯特, 和佩西·梁。从问答对中进行Freebase语义解析。在2013年经验自然语言处理会议论文集, 第1533–1544页, 2013年.[BDD⁺09] 路易莎·本蒂沃利, 伊多·达甘, 胡亚 Trang·达甘, 丹尼尔·吉安皮科洛, 和伯纳多·马格尼尼。第五届PASCAL文本蕴涵识别挑战。2009年.[BES10] 斯特凡诺·巴恰内拉, 安德烈亚·埃苏利, 和弗拉基泽奥·塞巴斯蒂亚尼。Sentiwordnet 3.0: 一个增强的用于情感分析和意见挖掘的词汇资源。在Lrec, 第10卷, 第2200–2204页, 2010年.[BHDD⁺06] 罗伊·巴·海姆, 伊多·达甘, 比尔·多兰, 丽莎·费罗, 丹尼尔·吉安皮科洛, 伯纳多·马格尼尼, 和伊丹·斯佩克托。第二届PASCAL识别文本蕴涵挑战。2006年.[BHT⁺20] 约纳坦·比斯克, 阿里·霍尔特曼, 杰西·托马斯, 雅各布·安德烈亚斯, 约书亚·本吉奥, 乔伊斯·蔡, 米雷拉·拉帕塔, 安杰利基·拉扎里杜, 乔纳森·梅, 亚历山大·尼斯涅维奇, 等。经验奠定了语言的基础。arXiv预印本 arXiv:2004.10151, 2020年.[BLC13] 约书亚·本吉奥, 尼古拉斯·L·leonard, 和亚伦·C·库里夫。通过随机神经元估计或传播梯度以进行条件计算。Arxiv, 2013年.[BZB⁺19] 约纳坦·比斯克, 罗温·泽勒斯, 罗南·勒·布拉, Jianfeng·高, 和叶珍·蔡。Piqa: 在自然语言中推理关于物理常识。arXiv预印本 arXiv:1911.11641, 2019年.[Car97] 理查德·卡鲁阿纳。多任务学习。机器学习, 28(1), 1997年.[CB78] 苏珊·凯里和埃尔莎·巴雷特。习得一个新词。斯坦福儿童语言会议论文集, 1978年.[CCE⁺18] 彼得·克拉克, 艾萨克·考黑, 奥伦·埃齐奥尼, 塔沙尔·科特, 阿什·萨巴瓦尔, 卡莉莎·肖尼克, 和奥伊文德·塔夫约德。你认为你解决了问答问题? 试试arc, ai2推理挑战。ArXiv, abs/1803.05457, 2018年.[CGRS19] 里温·查尔德, 斯科特·格雷, 亚历克·拉德福德, 和伊利亚·苏茨凯弗。使用稀疏Transformer生成序列, 2019年.[CHI⁺18] 恩索尔·蔡, 何, 莫希特·伊耶尔, 马克·亚茨卡, 文塔·伊, 叶珍·蔡, 佩西·梁, 和卢克·泽特勒莫耶。Quac: 上下文中的问答。Arxiv, 2018年.[CLC⁺19] 克里斯托弗·克拉克, 肯顿·李, 明伟·张, 汤姆·奎亚托斯基, 迈克尔·科林斯, 和克里斯蒂娜·塔特诺娃。BoolQ: 探索自然是否问题的惊人难度。arXiv预印本 arXiv:1905.10044, 2019年.[CLY⁺19] 尹纯·陈, 林洁·李, 李成, 艾哈迈德·埃尔·科利, 费萨尔·阿赫迈德, 甘·哲, 程·宇, 和刘·静静。Uniter: 学习通用图像-文本表示。arXiv预印本 arXiv:1909.11740, 2019年.[Cra17] 凯特·克劳福德。偏见的问题。NIPS 2017主题演讲, 2017年.[DCLT18] 雅各布·德夫林, 明伟·张, 肯顿·李, 和克里斯蒂娜·塔特诺娃。BERT: 为语言理解而预训练深度双向Transformer。arXiv预印本 arXiv:1810.04805, 2018年。

- [DGM06] Ido Dagan, Oren Glickman, and Bernardo Magnini. The PASCAL recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising textual entailment*, pages 177–190. Springer, 2006.
- [DGV⁺18] Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. Universal transformers. *Arxiv*, 2018.
- [DHKH14] Nadir Durrani, Barry Haddow, Philipp Koehn, and Kenneth Heafield. Edinburgh’s phrase-based machine translation systems for wmt-14. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 97–104, 2014.
- [DL15] Andrew M. Dai and Quoc V. Le. Semi-supervised sequence learning. In *Advances in neural information processing systems*, 2015.
- [DMST19] Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. The CommitmentBank: Investigating projection in naturally occurring discourse. 2019. To appear in proceedings of Sinn und Bedeutung 23. Data can be found at <https://github.com/mcdm/CommitmentBank/>.
- [DSC⁺16] Yan Duan, John Schulman, Xi Chen, Peter L. Bartlett, Ilya Sutskever, and Pieter Abbeel. RI²: Fast reinforcement learning via slow reinforcement learning. *Arxiv*, abs/1611.02779, 2016.
- [DWD⁺19] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *arXiv preprint arXiv:1903.00161*, 2019.
- [DYY⁺19] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *Arxiv*, 2019.
- [EOAG18] Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*, 2018.
- [FAL17] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *Arxiv*, abs/1703.03400, 2017.
- [Fyo00] Yaroslav Fyodorov. A natural logic inference system, 2000.
- [GG19] Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862*, 2019.
- [GLT⁺20] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*, 2020.
- [GMDD07] Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, pages 1–9. Association for Computational Linguistics, 2007.
- [Gra16] Alex Graves. Adaptive computation time for recurrent neural networks. *Arxiv*, 2016.
- [GSL⁺18] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*, 2018.
- [GSR19] Sebastian Gehrmann, Hendrik Strobelt, and Alexander M. Rush. Gltr: Statistical detection and 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.
- [HBFC19] Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *CoRR*, abs/1904.09751, 2019.
- [HLW⁺20] Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedziec, Rishabh Krishnan, and Dawn Song. Pretrained transformers improve out of distribution robustness. *arXiv preprint arXiv:2004.06100*, 2020.

- [DGM06] 伊利亚·苏茨凯弗、Oren Glickman 和 Bernardo Magnini. PASCAL 识别文本蕴涵挑战。在机器学习挑战：评估预测不确定性、视觉物体分类和识别文本蕴涵，第 177–190 页。Springer, 2006.[DGV⁺18] Mostafa Dehghani、Stephan Gouws、Oriol Vinyals、Jakob Uszkoreit 和 Lukasz Kaiser。通用 Transformer。Arxiv, 2018.[DHKH14] Nadir Durrani、Barry Haddow、Philipp Koehn 和 Kenneth Heafield。爱丁堡的基于短语的机器翻译系统用于 wmt-14。在第九届统计机器翻译研讨会论文集，第 97–104 页，2014.[DL15] Andrew M. Dai 和 Quoc V. Le。半监督序列学习。在神经信息处理系统进展，2015.[DMST19] Marie-Catherine De Marneffe、Mandy Simons 和 Judith Tonhauser。CommitmentBank：研究自然话语中的投射。2019。将发表在 Sinn und Bedeutung 23 研讨会论文集。数据可以在 <https://github.com/mcdm/CommitmentBank/> 找到.[DSC⁺16] Yan Duan、John Schulman、Xi Chen、Peter L. Bartlett、伊利亚·苏茨凯弗 和 Pieter Abbeel。RI²：通过慢速强化学习实现快速强化学习。Arxiv, abs/1611.02779, 2016.[DWD⁺19] Dheeru Dua、Yizhong Wang、Pradeep Dasigi、Gabriel Stanovsky、Sameer Singh 和 Matt Gardner。DROP：一个需要段落离散推理的阅读理解基准。arXiv 预印本 arXiv:1903.00161, 2019.[DYY⁺19] Zihang Dai、Zhilin Yang、Yiming Yang、Jaime G. Carbonell、Quoc V. Le 和 Ruslan Salakhutdinov。Transformer-xl：超出固定长度上下文的注意力语言模型。Arxiv, 2019.[EOAG18] Sergey Edunov、Myle Ott、Michael Auli 和 David Grangier。大规模理解回译。arXiv 预印本 arXiv:1808.09381, 2018.[FAL17] Chelsea Finn、Pieter Abbeel 和 Sergey Levine。模型无关元学习用于深度网络的快速适应。Arxiv, abs/1703.03400, 2017.[Fyo00] Yaroslav Fyodorov。一个自然逻辑推理系统，2000.[GG19] Hila Gonen 和 Yoav Goldberg。口红涂在猪上：去偏方法掩盖了词嵌入中的系统性性别偏见，但并未消除它们。arXiv 预印本 arXiv:1903.03862, 2019.[GLT⁺20] Kelvin Guu、Kenton Lee、Zora Tung、Panupong Pasupat 和 Ming-Wei Chang。Realm：检索增强的语言模型预训练。arXiv 预印本 arXiv:2002.08909, 2020.[GMDD07] Danilo Giampiccolo、Bernardo Magnini、伊利亚·苏茨凯弗 和 Bill Dolan。第三届 PASCAL 识别文本蕴涵挑战。在 ACL-PASCAL 研讨会论文集：文本蕴涵和释义，第 1–9 页。计算语言学协会，2007.[Gra16] Alex Graves。循环神经网络的适应计算时间。Arxiv, 2016.[GSL⁺18] Suchin Gururangan、Swabha Swayamdipta、Omer Levy、Roy Schwartz、Samuel R Bowman 和 Noah A Smith。自然语言推理数据中的标注伪影。arXiv 预印本 arXiv:1803.02324, 2018.[GSR19] Sebastian Gehrmann、Hendrik Strobelt 和 Alexander M. Rush。Gltr：生成文本的统计检测和可视化。arXiv 预印本 arXiv: 1906.04043, 2019.[GWC⁺18] Jiatao Gu、Yong Wang、Yun Chen、Kyunghyun Cho 和 Victor OK Li。低资源神经机器翻译的元学习。arXiv 预印本 arXiv:1808.08437, 2018.[HB20] Daniel Hernandez 和 Tom Brown。人工智能和效率，2020 年 5 月。[HBFC19] Ari Holtzman、Jan Buys、Maxwell Forbes 和 Yejin Choi。神经文本退化的奇怪情况。CoRR, abs/1904.09751, 2019.[HLW⁺20] Dan Hendrycks、Xiaoyuan Liu、Eric Wallace、Adam Dziedziec、Rishabh Krishnan 和 Dawn Song。预训练 Transformer 提高分布外鲁棒性。arXiv 预印本 arXiv:2004.06100, 2020。

- [HNA⁺17] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.
- [HR18] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018.
- [HVD15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
- [HYC01] Sepp Hochreiter, A Steven Younger, and Peter R Conwell. Learning to Learn Using Gradient Descent. In *International Conference on Artificial Neural Networks*, pages 87–94. Springer, 2001.
- [HZJ⁺19] Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. Reducing sentiment bias in language models via counterfactual evaluation. *arXiv preprint arXiv:1911.03064*, 2019.
- [IBGC⁺14] Mohit Iyyer, Jordan Boyd-Graber, Leonardo Claudino, Richard Socher, and Hal Daumé III. A neural network for factoid question answering over paragraphs. In *Empirical Methods in Natural Language Processing*, 2014.
- [IDCBE19] Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. Automatic detection of generated text is easiest when humans are fooled. *arXiv preprint arXiv:1911.00650*, 2019.
- [JCWZ17] Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.
- [JN20] Zheng Junyuan and Gamma Lab NYC. Numeric transformer - albert, March 2020.
- [JVS⁺16] Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*, 2016.
- [JYS⁺19] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. TinyBERT: Distilling BERT for natural language understanding. *arXiv preprint arXiv:1909.10351*, 2019.
- [JZC⁺19] Ying Ju, Fubang Zhao, Shijie Chen, Bowen Zheng, Xuefeng Yang, and Yunfeng Liu. Technical report on conversational question answering. *arXiv preprint arXiv:1909.10772*, 2019.
- [KCR⁺18] Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL)*, 2018.
- [KKS⁺20] Daniel Khashabi, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. Unifiedqa: Crossing format boundaries with a single qa system. *arXiv preprint arXiv:2005.00700*, 2020.
- [KMB20] Sarah E. Kreps, Miles McCain, and Miles Brundage. All the news that’s fit to fabricate: Ai-generated text as a tool of media misinformation, 2020.
- [KMH⁺20] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020.
- [KPR⁺19] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural 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.

[HNA⁺17] 乔尔·海斯特尼, 沙兰·纳兰, 纽什亚·阿达拉尼, 格雷戈里·戴亚莫斯, 希乌乌·俊, 哈桑·基安尼内贾德, Md.莫斯塔法·阿里·帕特瓦里, 杨杨, 和周燕琪。深度学习扩展是可预测的, 经验上。arXiv预印本 arXiv:1712.00409, 2017.[HR18] 杰里米·霍华德和塞巴斯蒂安·鲁德。通用语言模型微调用于文本分类。arXiv预印本 arXiv:1801.06146, 2018.[HVD15] 杰弗里·辛顿, 奥里奥尔·维尼亚尔斯, 杰夫·迪恩。神经网络中的知识蒸馏。arXiv预印本 arXiv:1503.02531, 2015.[HYC01] 塞普·霍赫莱特, A·史蒂文·扬格, 彼得·R·康韦尔。使用梯度下降学习。在人工智能神经网络国际会议, 第87-94页。斯普林格, 2001.[HZJ⁺19] 黄伯森, 张欢, 蒋雷, 罗伯特·斯坦福, 约翰内斯·韦尔布尔, 杰克·雷, 维沙尔·马尼, 达尼·约加塔玛, Pushmeet Kohli。通过反事实评估减少语言模型的情感偏差。arXiv预印本 arXiv:1911.03064, 2019.[IBGC⁺14] 莫希特·伊耶尔, 乔丹·博伊德-格雷伯, 莱昂纳多·克劳迪诺, 理查德·索彻, 哈尔·达姆·e III。用于段落事实性问答的神经网络。在自然语言处理经验方法, 2014.[IDCBE19] 达芙妮·伊波利托, 丹尼尔·达克沃斯, 克里斯·卡利斯昂·伯奇, 道格拉斯·埃克。自动检测生成文本在人类被愚弄时最容易。arXiv预印本 arXiv:1911.00650, 2019.[JCWZ17] 曼达尔·乔希, 刘恩索尔·赵, 丹尼尔·S·韦尔德, 卢克·泽特勒莫耶尔。TriviaQA: 一个大型的远程监督挑战数据集用于阅读理解。arXiv预印本 arXiv:1705.03551, 2017.[JN20] 郑俊元和Gamma Lab NYC。数字转换器 - albert, 2020年3月.[JVS⁺16] 拉法尔·乔泽福维茨, 奥里奥尔·维尼亚尔斯, 迈克·舒斯特, 诺亚·沙泽尔, 吴永辉。探索语言建模的极限。arXiv预印本 arXiv:1602.02410, 2016.[JYS⁺19] 焦晓奇, 尹一纯, 尚立峰, 蒋欣, 陈晓, 李琳琳, 王方, 刘群。TinyBERT: 为自然语言理解蒸馏BERT。arXiv预印本 arXiv:1909.10351, 2019.[JZC⁺19] 于莹, 赵福邦, 陈时杰, 郑 Bowen, 杨雪峰, 刘云峰。对话式问答技术报告。arXiv预印本 arXiv:1909.10772, 2019.[KCR⁺18] 丹尼尔·哈希abi, Snigdha Chaturvedi, 迈克尔·罗特, Shyam Upadhyay, 丹·罗特。超越表面: 一个用于多句阅读理解的挑战集。在北美计算语言学协会分会会议记录, 2018.[KKS⁺20] 丹尼尔·哈希abi, Tushar Khot, 阿希什·萨巴瓦尔, Oyvind Tafjord, 彼得·克拉克, Hannaneh Hajishirzi。Unifiedqa: 用一个问答系统跨越格式边界。arXiv预印本 arXiv:2005.00700, 2020.[KMB20] 莎拉·E·克雷普斯, 迈尔斯·麦肯尼, 迈尔斯·布兰达奇。所有适合纺织的新闻: AI生成的文本作为媒体错误信息的工具, 2020.[KMH⁺20] Jared Kaplan, 山姆·麦卡迪尔, 汤姆·亨尼汉, Tom B. Brown, 本杰明·切斯, 雷温·查尔德, 斯科特·格雷, 亚历克·拉德福德, 杰弗里·吴, 达里奥·阿莫迪。神经语言模型的规模定律, 2020.[KPR⁺19] 汤姆·克维亚托夫斯基, Jennimaria Palomaki, 奥利维亚·雷德菲尔德, 迈克尔·科林斯, 安库尔·帕里赫, 克里斯·阿尔贝蒂, 丹尼尔·爱泼斯坦, 伊利亚·波洛苏金, 马修·凯尔西, 雅各布·德夫林, 肯顿·李, 克里斯蒂娜·N·塔诺瓦娜, 利昂·琼斯, 明·伟·张, 安德鲁·戴, 雅各布·乌斯克雷特, Quoc Le, 斯拉夫·佩特罗夫。自然问题: 一个问题回答研究的基准。计算语言学协会交易, 2019.[KR16] 尹·金和亚历山大·M·鲁什。序列级知识蒸馏。Arxiv, 2016.[LB02] 爱德华·洛珀和史蒂文·伯德。Nltk: 自然语言工具包, 2002.[LC19] 吉约姆·兰普勒和阿莱克斯·孔诺。跨语言模型预训练。arXiv预印本 arXiv:1901.07291, 2019。

- [LCG⁺19] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricu. 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.
- [LDL19] Zhongyang Li, Xiao Ding, and Ting Liu. Story ending prediction by transferable bert. *arXiv preprint arXiv:1905.07504*, 2019.
- [LDM12] Hector Levesque, Ernest Davis, and Leora Morgenstern. The Winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*, 2012.
- [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.
- [LH17] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [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.
- [Lin20] Tal Linzen. How can we accelerate progress towards human-like linguistic generalization? *arXiv preprint arXiv:2005.00955*, 2020.
- [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.
- [LSP⁺18] Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. Generating Wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*, 2018.
- [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.
- [LXL⁺17] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*, 2017.
- [LYN⁺20] Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and Jimmy Lin. Ttttackling winogrande schemas. *arXiv preprint arXiv:2003.08380*, 2020.
- [Mac92] David. MacKay. Information-based objective functions for active data selection. *Neural Computation*, 1992.

- [LCG⁺19] 兰振中, 陈明达, 塞巴斯蒂安·古德曼, 凯文·金佩尔, 皮尤什·夏尔马, 和拉杜·索里-库特。ALBERT: 一个用于自监督学习语言表示的轻量级BERT。arXiv预印本arXiv:1909.11942, 2019.[LCH⁺20] 刘晓东, 程浩, 何鹏程, 陈伟珠, 王宇, 潘海锋, 和高建锋。大型神经语言模型的对抗训练。arXiv预印本arXiv:2004.08994, 2020.[LDL19] 李忠, 丁晓, 和刘婷。通过可迁移BERT进行故事结尾预测。arXiv预印本arXiv:1905.07504, 2019.[LDM12] 赫克托·莱夫斯克, 埃尔南·戴维斯, 和莱奥拉·莫根施特恩。Winograd Schema挑战。在第十三届知识表示与推理原理国际会议, 2012.[LGG⁺20] 刘银汉, 顾嘉涛, 高南, 李翔, 谢尔盖·艾杜诺夫, 马贾恩·加兹维尼内贾德, 迈克·刘易斯, 和卢克·泽特勒莫耶。神经机器翻译的多语言去噪预训练。arXiv预印本arXiv:2001.08210, 2020.[LGH⁺15] 刘晓东, 高建锋, 何晓东, 邓力, 杜凯文, 和王叶怡。用于语义分类和信息检索的多任务深度神经网络表示学习。在2015年北美计算语言学协会会议: 人机语言技术, 2015.[LH17]伊利亚·洛斯基洛夫和弗兰克·胡特。解耦权重衰减正则化。arXiv预印本arXiv:1711.05101, 2017.[LHCG19a] 刘晓东, 何鹏程, 陈伟珠, 和高建锋。通过知识蒸馏改进多任务深度神经网络以用于自然语言理解。arXiv预印本arXiv:1904.09482, 2019.[LHCG19b] 刘晓东, 何鹏程, 陈伟珠, 和高建锋。用于自然语言理解的多任务深度神经网络。arXiv预印本arXiv:1901.11504, 2019.[Lin20] 塔尔·林森。我们如何才能加速实现类人语言泛化的进展? arXiv预印本arXiv:2005.00955, 2020.[LLG⁺19] 迈克·刘易斯, 刘银汉, 高南, 马贾恩·加兹维尼内贾德, 阿布杜勒拉赫曼·穆罕默德, 奥默·利维, 韦斯·斯托亚诺夫, 和卢克·泽特勒莫耶。BART: 用于自然语言生成、翻译和理解的去噪序列到序列预训练。arXiv预印本arXiv:1910.13461, 2019.[LM17] 李克和贾廷德拉·马利克。学习优化神经网络。arXiv预印本arXiv:1703.00441, 2017.[LOG⁺19] 刘银汉, 米莱·奥特, 高南, 杜景飞, 乔希·曼达尔, 陈丹琪, 奥默·利维, 迈克·刘易斯, 卢克·泽特勒莫耶, 和韦塞林·斯托亚诺夫。RoBERTa: 一种鲁棒优化的BERT预训练方法。arXiv预印本arXiv:1907.11692, 2019.[LPP⁺20] 帕特里克·刘易斯, 埃than·佩雷斯, 亚历山德拉·皮库斯, Fabio·佩特罗尼, 弗拉基米尔·卡普钦, 高南, 海因里希·库特勒, 迈克·刘易斯, 尹tau·伊, Tim·罗克塔施尔, 塞巴斯蒂安·里德尔, 和基拉·多夫。用于知识密集型NLP任务的检索增强生成。arXiv预印本arXiv:2005.11401, 2020.[LSP⁺18] 彼得·J·刘, 穆罕默德·萨利赫, 埃蒂安·波, 本·古德里奇, 瑞安·塞帕西, 卢卡兹·凯撒, 和诺亚姆·沙泽尔。通过总结长序列生成维基百科。arXiv预印本arXiv:1801.10198, 2018.[LWS⁺20] 李卓航, Eric·瓦莱尔, 沈胜, 凯文·林, 库尔特·基特泽, 丹·克莱因, 和约瑟夫·E·冈萨雷斯。先训练大型模型, 再压缩: 重新思考Transformer高效训练和推理的模型大小, 2020.[LXL⁺17] 赖国昆, 谢启哲, 刘汉晓, 杨毅明, 和Eduard·霍维。RACE: 来自考试的规模巨大的阅读理解数据集。arXiv预印本arXiv:1704.04683, 2017.[LYN⁺20] 林胜杰, 杨振宏, 罗德里戈·诺格利亚, 蔡明锋, 王传举, 和Jimmy·林。Ttttackling Winogrande Schema。arXiv预印本arXiv:2003.08380, 2020.[Mac92] 大卫·麦凯。基于信息的主动数据选择的目标函数。神经计算, 1992。

- [MBXS17] Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems*, pages 6294–6305, 2017.
- [MCCD13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
- [MCH⁺16] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and evaluation framework for deeper understanding of commonsense stories. *arXiv preprint arXiv:1604.01696*, 2016.
- [MCKS18] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. *ArXiv*, abs/1809.02789, 2018.
- [MKAT18] Sam McCandlish, Jared Kaplan, Dario Amodei, and OpenAI Dota Team. An empirical model of large-batch training, 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.
- [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.
- [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*, 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.
- [NBR20] Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint 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.
- [Nor09] Peter Norvig. Natural language corpus data, 2009.
- [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.
- [NWD⁺19] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*, 2019.
- [oR16] University of Regensburg. Fascha, 2016.
- [PCC18] Mohammad Taher Pilehvar and Jose Camacho-Collados. WIC: 10,000 example pairs for evaluating context-sensitive representations. *arXiv preprint arXiv:1808.09121*, 2018.
- [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.
- [PHR⁺18] Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. Collecting diverse natural language inference problems for sentence representation evaluation. In *Proceedings of EMNLP*, 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.
- [PNZtY18] Matthew E. Peters, Mark Neumann, Luke Zettlemoyer, and Wen tau Yih. Dissecting contextual word embeddings: Architecture and representation, 2018.
- [Pos18] Matt Post. A call for clarity in reporting BLEU scores. *arXiv preprint arXiv:1804.08771*, 2018.

[MBXS17] 布莱恩·麦卡恩, 詹姆斯·布拉德伯里, 谢民, 以及理查德·索彻。从翻译中学习: 上下文词向量。在《神经信息处理系统进展》中, 第 6294–6305 页, 2017 年。[MCCD13] 托马什·米科洛夫, 凯伊·陈, 格雷格·科拉德, 以及杰弗里·迪恩。向量空间中词表示的有效估计。arXiv 预印本 arXiv:1301.3781, 2013 年。[MCH⁺16] 纳斯林·莫斯塔法扎德, 纳撒尼尔·钱伯斯, 何晓东, 德维·帕里克, 德鲁夫·巴特拉, 露西·范德温德, 普什梅特·科利, 以及詹姆斯·艾伦。常识故事更深层理解的语料库和评估框架。arXiv 预印本 arXiv:1604.01696, 2016 年。[MCKS18] 托多尔·米哈伊洛夫, 彼得·克拉克, 图沙尔·科特, 以及阿希什·萨巴瓦尔。一套盔甲能导电吗? 一个用于开放式问答的新数据集。ArXiv, abs/1809.02789, 2018 年。[MKAT18] 山姆·麦卡尼斯利什, Jared Kaplan, 达里奥·阿莫迪, 以及 OpenAI Dota 队。大批量训练的实证模型, 2018 年。[MKM⁺94] 米切尔·马库斯, 恩典·金, 玛丽·安·马库斯凯维奇, 罗伯特·麦金泰尔, 安·比斯, 马克·弗格森, 卡伦·卡茨, 以及布丽塔·沙斯伯格。宾夕法尼亚树库: 标注谓词论元结构。在《人机语言技术研讨会论文集》中, 第 114–119 页。计算语言学协会, 1994 年。[MKXS18] 布莱恩·麦卡恩, 尼蒂什·希里什·凯斯卡尔, 谢民, 以及理查德·索彻。自然语言十项全能: 多任务学习作为问答。arXiv 预印本 arXiv:1806.08730, 2018 年。[MPL19] R 托马斯·麦科伊, 艾莉·帕夫利克, 以及塔尔·林岑。为了错误的原因: 诊断自然语言推理中的句法启发式。arXiv 预印本 arXiv:1902.01007, 2019 年。[MWZ⁺18] 玛格丽特·米切尔, 西蒙娜·吴, 安德鲁·扎尔迪瓦尔, 帕克·巴恩斯, 露西·瓦瑟曼, 本·哈钦森, 埃琳娜·斯皮策, 伊诺卢瓦·德博拉·拉吉, 以及蒂姆尼特·杰布鲁。模型卡用于模型报告, 2018 年。[NBR20] 毛因·纳迪姆, 安娜·贝克, 以及西瓦·雷迪。立体集: 测量预训练语言模型中的刻板印象偏差。arXiv 预印本 arXiv:2004.09456, 2020 年。[NK19] 蒂莫西·尼文和 Hung-Yu 高。探测神经网络对自然语言论点的理解。arXiv 预印本 arXiv:1907.07355, 2019 年。[Nor09] 彼得·诺维格。自然语言语料库数据, 2009 年。[NvNvdG19] 马尔文娜·尼西姆, 里克·范诺德, 以及罗布·范德古特。公平比耸人听闻更好: 男人对医生如女人对医生。arXiv 预印本 arXiv:1905.09866, 2019 年。[NWD⁺19] 聂依欣, 阿迪娜·威廉姆斯, 艾米丽·迪南, 莫希特·班萨尔, 杰森·韦斯顿, 以及多韦·基拉。对抗性 NLI: 自然语言理解的新基准。arXiv 预印本 arXiv:1910.14599, 2019 年。[oR16] 雷根斯堡大学。Fascha, 2016 年。[PCC18] 穆罕默德·塔赫尔·皮莱赫瓦尔和何塞·卡马乔·科利亚多斯。WiC: 10,000 个示例对用于评估上下文敏感表示。arXiv 预印本 arXiv:1808.09121, 2018 年。[PFB18] 杰森·方, 蒂博·费夫里, 以及塞缪尔·R·鲍曼。STILTs 上的句子编码器: 在中间标记数据任务上进行补充训练。arXiv 预印本 arXiv:1811.01088, 2018 年。[PHR⁺18] 亚当·波利亚克, 阿帕拉吉塔·哈尔达尔, 拉切尔·鲁丁格, J. 爱德华·胡, 艾莉·帕夫利克, 亚伦·史蒂文·怀特, 以及本杰明·范·杜尔姆。为句子表示评估收集多样化的自然语言推理问题。在 EMNLP 论文中, 2018 年。[PKL⁺16] 丹尼斯·帕佩诺, 杰尔曼·库尔谢夫斯基, 安吉莉克·拉扎里多乌, 阮秋玉·范, 拉法埃拉·贝尔纳迪, 桑德罗·佩泽莱, 马科·巴罗尼, 杰玛·博莱达, 以及拉奎尔·费尔南德斯。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 年。[PNZtY18] 马修·E·彼得斯, 马克·纽曼, 卢克·泽特莫耶尔, 以及魏泰·伊。剖析上下文词嵌入: 架构和表示, 2018 年。[Pos18] 马特·波斯特。呼吁在报告 BLEU 分数时保持清晰。arXiv 预印本 arXiv:1804.08771, 2018 年。

- [PSM14] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014.
- [QIA20] QIANXIN. Sa-net on albert (ensemble), April 2020.
- [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.
- [RBG11] Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*, 2011.
- [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.
- [RCP⁺17] Scott Reed, Yutian Chen, Thomas Paine, Aaron 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.
- [RJL18] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don’t know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*, 2018.
- [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.
- [RNSS18] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training, 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.
- [SBBC19] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale, 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.

[PSM14] 亚历克·彭宁顿、理查德·索彻和克里斯托弗·马恩。GloVe: 用于词表示的全局向量。在2014年自然语言处理经验方法会议 (EMNLP) 论文集中, 2014年.[QIA20] QIANXIN. Sa-net on albert (集成), 2020年4月.[QMZH19] 钱宇苏、乌尔瓦·穆阿兹、张本和吴宰温·Hyun。使用性别平等化损失函数减少词级语言模型的性别偏见。arXiv预印本 arXiv:1905.12801, 2019年.[RBG11] 梅丽莎·罗梅勒、科斯明·阿德里安·贝扬和安德鲁·S·戈登。合理替代的选择: 常识因果关系推理的评估。在2011年 AAAI 春季研讨会系列中, 2011年.[RCM19] Siva Reddy、陈丹琪和克里斯托弗·D·马恩。Coqa: 一个对话式问答挑战。计算语言学协会汇刊, 7:249–266, 2019年.[RCP⁺17] 斯科特·里德、陈宇天、托马斯·佩恩、Aaron van den Oord、SM Eslami、丹尼尔·雷泽德、奥里奥尔·维尼亚尔斯和南多·德·弗里塔斯。少样本自回归密度估计: 朝着学习如何学习分布的方向。arXiv预印本 arXiv:1710.10304, 2017年.[RJL18] 普拉纳夫·拉贾普卡尔、Jia Robin和Percy Liang。知道你不知道什么: 为squad的无答案问题。arXiv预印本 arXiv:1806.03822, 2018年.[RL16] Sachin Ravi和Hugo Larochelle。优化作为少样本学习的一个模型。ICLR 2017 (口头), 2016年.[RLL⁺19] 钱求、林彦开、李鹏、周杰和刘志远。NumNet: 使用数值推理的机器阅读理解。在EMNLP论文集中, 2019年.[RNLVD18] Rachel Rudinger、Jason Naradowsky、Brian Leonard和 Benjamin Van Durme。核心指代中的性别偏见。arXiv预印本 arXiv:1804.09301, 2018年.[RNSS18] 亚历克·拉德福德、卡齐克·纳拉辛汉、Tim Salimans和伊利亚·苏茨凯弗。通过生成式预训练提高语言理解, 2018年.[Ros12] R.S. Ross。进行风险评估的指南。NIST特别出版物, 2012年.[RRBS19] 乔纳森·S·罗森菲尔德、阿米尔·罗森菲尔德、Yonatan Belinkov和Nir Shavit。跨尺度的泛化误差的构造性预测, 2019年.[RRS20] 亚当·罗伯茨、科林·拉夫尔 and Noam Shazeer。你能把多少知识塞进语言模型的参数中? arXiv预印本 arXiv:2002.08910, 2020年.[RSR⁺19] 科林·拉夫尔、Noam Shazeer、亚当·罗伯茨、凯瑟琳·李、Sharan Narang、Michael Matena、周燕琪、李伟和Peter J. Liu。探索迁移学习的极限, 使用一个统一的文本到文本转换器, 2019年.[RWC⁺19] 亚历克·拉德福德、杰弗里·吴、雷温·查尔德、David Luan、达里奥·阿莫迪和伊利亚·苏茨凯弗。语言模型是无监督的多任务学习器, 2019年.[SBBC19] Sakaguchi Keisuke、Ronan Le Bras、Chandra Bhagavatula和Yejin Choi。Winogrande: 大规模的对抗性 Winograd Schema挑战, 2019年.[SBC⁺19] Irene Solaiman、Miles Brundage、Jack Clark、Amanda Askell、Ariel Herbert-Voss、Jeff Wu、亚历克·拉德福德、格雷琴·克鲁格、Jong Wook Kim、Sarah Kreps、Miles McCain、Alex Newhouse、Jason Blazakis、Kris McGuffie和Jasmine Wang。发布策略和语言模型的社会影响, 2019年.[SCNP19] Emily Sheng、Kai-Wei Chang、Premkumar Natarajan和Nanyun Peng。女人当过保姆: 关于语言生成的偏见。arXiv预印本 arXiv:1909.01326, 2019年.[SDCW19] Victor Sanh、Lysandre Debut、Julien Chaumond和Thomas Wolf。DistilBERT, BERT的提炼版本: 更小、更快、更便宜、更轻。arXiv预印本 arXiv:1910.01108, 2019年.[SDSE19] Roy Schwartz、Jesse Dodge、Noah A. Smith和Oren Etzioni。绿色人工智能。CoRR, abs/1907.10597, 2019年.[SHB15] Rico Sennrich、Barry Haddow和Alexandra Birch。使用单语数据改进神经机器翻译模型。arXiv预印本 arXiv:1511.06709, 2015年。

- [SMM⁺17] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarczyk, 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.
- [TFR⁺17] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 23–30. IEEE, 2017.
- [TL05] Peter D. Turney and Michael L. Littman. Corpus-based learning of analogies and semantic relations. *CoRR*, abs/cs/0508103, 2005.
- [TL18] Trieu H. Trinh and Quoc V. Le. A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*, 2018.
- [TLBS03] Peter D. Turney, Michael L. Littman, Jeffrey Bigham, and Victor Shnayder. Combining independent modules to solve multiple-choice synonym and analogy problems. *CoRR*, cs.CL/0309035, 2003.
- [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.
- [YdC⁺19] Dani Yogatama, Cyprien de Masson d’Autume, Jerome Connor, Tomas Kocisky, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, et al. Learning and evaluating general linguistic intelligence. *arXiv preprint arXiv:1901.11373*, 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.
- [ZLL⁺18] Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. ReCoRD: Bridging the gap between human and machine commonsense reading comprehension. *arXiv preprint arXiv:1810.12885*, 2018.
- [ZSW⁺19a] 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.

- [SMM⁺17] 诺亚·沙泽尔、阿扎莉亚·米尔霍塞尼、克日什托夫·马兹亚尔兹、安迪·戴维斯、Quoc Le、杰弗里·辛顿和杰夫·迪恩。 outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.[SPP⁺19] 穆罕默德·肖伊比、莫斯塔法·帕特瓦里、劳尔·普里、帕特里克·勒格雷西、贾雷德·卡斯珀和Bryan Catanzaro. Megatron-lm: 使用模型并行训练数十亿参数的语言模型, 2019.[SS20] 蒂莫·希克和Hinrich Schütze. 利用完形填空问题进行少样本文本分类和自然语言推理. *arXiv preprint arXiv:2001.07676*, 2020.[STQ⁺19] 宋凯涛、Tan Xu、秦涛、陆建峰和刘提炎。 MASS: 用于语言生成的掩码序列到序列预训练. *arXiv preprint arXiv:1905.02450*, 2019.[TFR⁺17] 乔什·托宾、Rachel Fong、Alex Ray、Jonas Schneider、Wojciech Zaremba和Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 23–30. IEEE, 2017.[TL05] 彼得·D·特尼和迈克尔·L·利特曼。 Corpus-based learning of analogies and semantic relations. *CoRR*, abs/cs/0508103, 2005.[TL18] Trieu H. Trinh和Quoc V. Le. A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*, 2018.[TLBS03] 彼得·D·特尼、迈克尔·L·利特曼、Jeffrey Bigham和Victor Shnayder. Combining independent modules to solve multiple-choice synonym and analogy problems. *CoRR*, cs.CL/0309035, 2003.[Tur20] Project Turing. 微软研究博客, 2020年2月.[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] 阿希什·瓦桑维、诺亚·沙泽尔、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N. Gomez、Łukasz Kaiser和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和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] 王一仁、夏英采、何天宇、田飞、秦涛、翟成祥和张铁岩。 Multi-agent dual learning. *ICLR 2019*, 2018.[XDH⁺19] 谢启哲、戴志航、Eduard Hovy、Minh-Thang Luong和Quoc V. Le. Unsupervised data augmentation for consistency training, 2019.[YdC⁺19] Dani Yogatama、Cyprien de Masson d’Autume、杰罗姆·康纳、Tomas Kocisky、Mike Chrzanowski、Lingpeng Kong、Angeliki Lazaridou、王玲、刘雷、Chris Dyer, et al. Learning and evaluating general linguistic intelligence. *arXiv preprint arXiv:1901.11373*, 2019.[YDY⁺19] 杨志林、戴志航、杨毅明、杰伊·卡本内尔、Ruslan Salakhutdinov和Quoc V. Le. XLNet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*, 2019.[ZHB⁺19] 罗恩·泽勒斯、Ari Holtzman、Yonatan Bisk、Ali Farhadi和Yejin Choi. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.[ZHR⁺19] 罗恩·泽勒斯、Ari Holtzman、Hannah Rashkin、Yonatan Bisk、Ali Farhadi、Franziska Roesner和Yejin Choi. Defending against neural fake news. *arXiv preprint arXiv:1905.12616*, 2019.[ZLL⁺18] 张胜、刘东东、刘静静、高建峰、Kevin Duh和Benjamin Van Durme. ReCoRD: Bridging the gap between human and machine commonsense reading comprehension. *arXiv preprint arXiv:1810.12885*, 2018.[ZSW⁺19a] 丹尼尔·M·齐格勒、Nisan Stiennon、Jeffrey Wu、Tom B. Brown、亚历克·拉德福德、达里奥·阿莫迪、Paul Christiano和杰弗里·欧文。 Fine-tuning language models from human preferences, 2019.

[ZSW⁺19b] 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. *ArXiv*, abs/1909.08593, 2019.

[ZSW⁺19b] 丹尼尔·M·齐格勒、尼桑·斯蒂尼昂、杰弗里·吴、Tom B. Brown、亚历克·拉德福德、达里奥·阿莫迪、保罗·克里斯蒂安诺和杰弗里·欧文。从人类偏好中微调语言模型。 *ArXiv*, abs/1909.08593, 2019。