Glam：具有专家混合物的语言模型的有效尺度

Nan Du [*1] **Yanping Huang** [*1] **Andrew M. Dai** [*1] **Simon Tong** [1] **Dmitry Lepikhin** [1] **Yuanzhong Xu** [1] **Maxim Krikun** [1] **Yanqi Zhou** [1] **Adams Wei Yu** [1] **Orhan Firat** [1] **Barret Zoph** [1] **Liam Fedus** [1] **Maarten Bosma** [1] **Zongwei Zhou** [1] **Tao Wang** [1] **Yu Emma Wang** [1] **Kellie Webster** [1] **Marie Pellat** [1] **Kevin Robinson** [1] **Kathleen Meier-Hellstern** [1] **Toju Duke** [1] **Lucas Dixon** [1] **Kun Zhang** [1] **Quoc V Le** [1] **Yonghui Wu** [1] **Zhifeng Chen** [1] **Claire Cui** [1]

Abstract

使用更多数据缩放语言模型，ComputeAnd参数驱动了进步的重要进步内在语言处理。例如，感谢缩放，GPT-3能够实现对文化学习任务的强有力重新塑造。但是，培训这些大型密集的模型需要大量的计算资源。在本文中，我们提出并开发了一个名为Glam（通才语言模型）的语言模型家族，该语言模型（通才语言模型）使用稀疏激活的ExpertSharchituction来扩展模型容量，同时比相比的培训成本少于密集的变体。最大的魅力具有1.2个三角体参数，大约为7倍GPT-3。它仅消耗1/3的能源来训练GPT-3，并且需要一半的com-point频率进行推理，而仍达到总体零，一零和几乎没有射击的29个NLP任务。

一、简介

语言模型在过去截然不同的自然语言处理（NLP）中发挥了重要作用。语言模型的变体已被用来培训预验证的单词向量（Mikolov等，2013; Penning-Ton等，2014）和上下文化的单词矢量（Peterset al。，2018; Devlin等，2019）对于许多NLP应用程序。使用更多数据和更大的Mod-els进行扩展的转变（Shazeer等，2017； Huang等，2019; Kaplan等，2020），使复杂的自然语言任务能够按标记的数据较少。例如，GPT-3（Brownet Al。，2020）和Flan（Wei等，2021）证明了在几次射击甚至零射的概括中学习的可行性，这意味着很少有标记的示例为在NLP应用程序上取得良好的性能。虽然有效且性能进一步缩放量很高，而且更加昂贵，并且能够消耗明显的能量。（Patterson等，2021）。

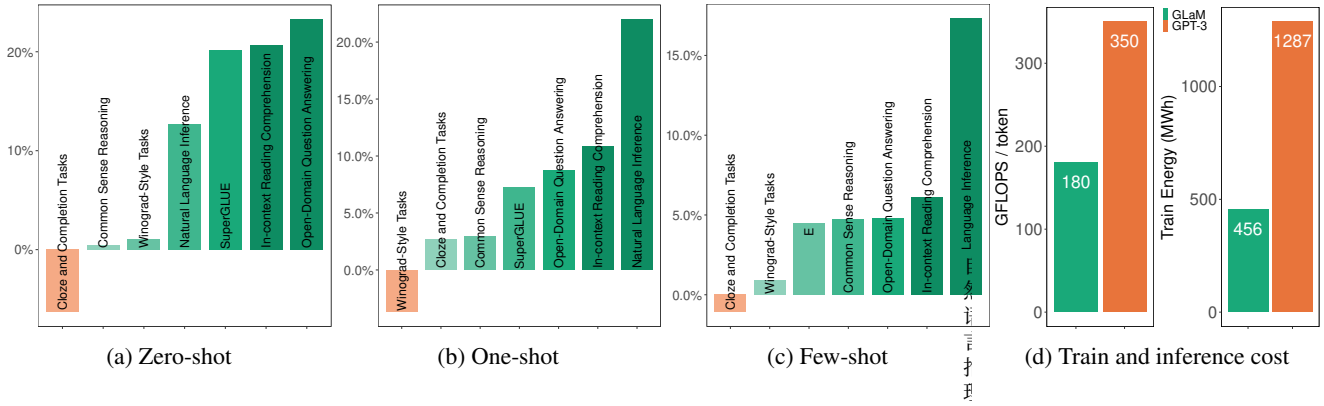
在这项工作中，我们表明，与最新的少数任务上的最新模型相比，大量稀疏激活的NetworkCan取得了竞争成果，同时更加有效。我们提出了一个称为Glam的通才局域网模型的家族，该模型触动了余额和有条件的计算。 Glam的最大版本总共具有1.2T参数，有64个专家permoe层（Shazeer等，2017； Lepikhin等，2021; Fe-Dus等，2021），在其中，每个在输入批处理中的每个值都只能激活一个子网络。 96.6b（占1.2t）参数的零，一个且很少的学习，该模型可与GPT-3（175b）进行比较，并在29个公共NLP基准中进行了显着改进的效率，从语言完成任务范围内，开放域QA任务，自然语言推理任务。得益于稀疏活化的体系结构以及对theDel并行性算法的有效实施，总能量消耗培训仅占GPT-3的三分之一。我们强调了表1中最大版本的Glam andGPT-3和图1之间的比较。

表1。GPT-3和Glam之间的比较。简而言之，Glam在21种自然语言理解（NLU）基准（NLU）基准和8种自然语言生成（NLG）基准的平均胜过GPT-3的同时，使用大约一半的flops flops，每个flops flops tokends tokend tokend tokent tokent tokende tokent tokend tokent tokende tokende tokends tokends interpects andertians tokends interpects and to tokends interpects and ofter and to the tokent intermands和3。

		GPT-3	GLaM	relative
cost	FLOPs / token (G)	350	**180**	**–48.6%**
	Train energy (MWh)	1287	**456**	**–64.6%**
accuracy on average	Zero-shot	56.9	**62.7**	**+10.2%**
	One-shot	61.6	**65.5**	**+6.3%**
	Few-shot	65.2	**68.1**	**+4.4%**

*同等贡献1 Google。通讯：Nan Du, Yanping Huang和Andrew M. Dai 。

第 39 届国际机器学习会议论文集，美国马里兰州巴尔的摩，PMLR 162，2022 年。



(a) Zero-shot (b) One-shot (c) Few-shot (d) Train and inference cost

图1。概述了GLAM（64b/64e）的预测性能变化百分比变化（较高）与（a）零击中的GPT-3（175b）的概述，（b）一击和（c）在7个基准类别中，总共有29个公共任务的7个基准设置。面板（a），（b）和（c）中的每个栏都代表一个基准类别。面板（D）比较了每个令牌预测和训练能量消费所需的拖船。

我们使用Glam研究数据的重要性。我们的分析方法是，即使对于这些大型模型，如果目标是产生高质量的自动回归语言模型，那么数据质量也不应牺牲数量。更重要的是，在社会维度上，我们的结果也是我们的知识的首先要缩小WinogenderBenchmark上立体类典型和反性典型示例之间的性能差距，这表明大型，稀疏激活的模型依赖于超级统计相关性。

最后，尽管基于MOE的稀疏模型尚未在NLP社区中使用，但我们的作品表明，仅SparsEdeCoder语言模型可以比第一次在几次爆发中的第一次使用类似的计算机拖鞋的密集体系结构更具性能学习设置ATSCALE，表明稀疏性是实现高质量NLP模型同时节省效果成本的最有希望的导向之一（Patterson等，2021）。因此，MOE应该被视为未来扩展的有力候选人。

2.相关工作

语言模型。神经语言模型（Mikolovet al。，2010; Sutskever等，2011）已被证明对许多自然语言处理任务感到满意。 Word2Vec（Mikolovet al。，2013），手套（Pennington等人，2014年）和段落向量（Le＆Mikolov，2014年）等单词Em床单模型和扩展。

预训练和微调。丰富的COM-PUTE和数据使越来越大的模型可以通过预先培训的预训练。对于训练通知网络表现出显着的可伸缩性，这是自然拟合。使用诸如RNN和LSTMS forlanguage表示等复发模型的工作元（Dai＆Le，2015; Kiros等，2015）表明，通用语言模型可以拟合

改善各种语言理解任务。更重复地，使用变压器的模型（Vaswani等人，2017年）表明，在不贝型数据上进行自学的较大模型可以对NLPTASKS产生显着改进（Devlin等，2019; Yang等，2019，2019 ; Liu等人，2019年； Clark等，2020）。基于培训前和填充的转移学习（Raffel等，2020; Houlsby等，2019）已被广泛研究，并在下游任务上表现出了很好的表现。但是，这种方法的主要限制是它需要任务特定的调整。

在文化中几次学习。 GPT-3（Brown等，2020）和相关工作（Shoeybi等，2019; Lieber等，2021; Wei等，2021）表明，扩展Lan-Guage模型可以大大改善任务无关，每种形式的镜头很少。这些语言模型是无需任何gr缩更新而应用的，只有少数通过与模型的文本交互纯粹针对的演示。

稀疏的封闭式网络。基于专家的混合物也显示出显着的优势。对于Lan-Guage建模和机器翻译，Shazeer等人（2017年）表明，它们可以有效地使用非常额外的权重，而只需要在推理时间计算计算图的小材料。因此，还在研究稀疏激活的Moe Ar-Chitectures（Hestness等，2017; Shazeer等，2018; Lep-ikhin等，2021; Kudugunta et al。，2021）。最近，Feduset Al。（2021）显示了更大的1万亿pa型符合物稀疏激活模型（Switch-C）的结果。尽管Bloth Switch-C和最大的GLAM模型具有一个三角数的可训练参数，但Glam是一个仅限训练的语言模型，而Switch-C是基于编码的序列序列的序列模型。此外，Switch-C主要在填充基准测试中进行评估，例如Superglue，而Glam的性能很好，没有任何

e.g., SuperGlue, while GLaM performs well without any

表2。相关模型的样本（Devlin等，2019; Rafflet al。，2020; Brown等，2020; Lieber等，2021; Rae等，2021; Shoeybi等，2019; Lepikhin等人，2021年； NPARAMS是可吸引的模型参数的总数，NACT-PARAMS是每个输入令牌的激活模型参数的数量。

Model Name	Model Type	n_{params}	$n_{\text{act-params}}$
BERT	Dense Encoder-only	340M	340M
T5	Dense Encoder-decoder	13B	13B
GPT-3	Dense Decoder-only	175B	175B
Jurassic-1	Dense Decoder-only	178B	178B
Gopher	Dense Decoder-only	280B	280B
Megatron-530B	Dense Decoder-only	530B	530B
GShard-M4	MoE Encoder-decoder	600B	1.5B
Switch-C	MoE Encoder-decoder	1.5T	1.5B
GLaM (64B/64E)	MoE Decoder-only	1.2T	96.6B

GPT-3 Where Superglue共享的几弹性设置中进行填充的需求是子集。表2总结了Glam和相关模型前培训文本语料库之间的键差。

3。培训数据集

为了培训我们的模型，我们构建了一个16亿代表的高质量数据集，这些数据集代表了广泛的语言用例。网页构成了我们未标记的数据集中数据的广阔性。但是，他们的质量范围从专业写作到低质量的委员会和论坛页面。与Brown等人类似。（2020），与我们自己的文本质量分类器相关，以从原始的较大的原始语料库中产生高质量的网络语料库。 WEUSE基于功能的基于哈希的线性分类器的推理速度。该分类器经过培训，可以在策划的文本（Wikipedia，Books和一些选定的Web网站）和其他网页之间进行分类。我们使用此分类器来估计网页的内容质量。然后，我们通过使用Pareto分布来将此CLAS-SIFIER应用于其得分来采样WebPagesAccording。这允许包括一些低质量网络，以防止群体中的系统偏见（Brown等，2020）。

表3。华丽训练集中的数据和混合物。

Dataset	Tokens (B)	Weight in mixture
Filtered Webpages	143	0.42
Wikipedia	3	0.06
Conversations	174	0.28
Forums	247	0.02
Books	390	0.20
News	650	0.02

我们使用此过程来生成高质量过滤子集

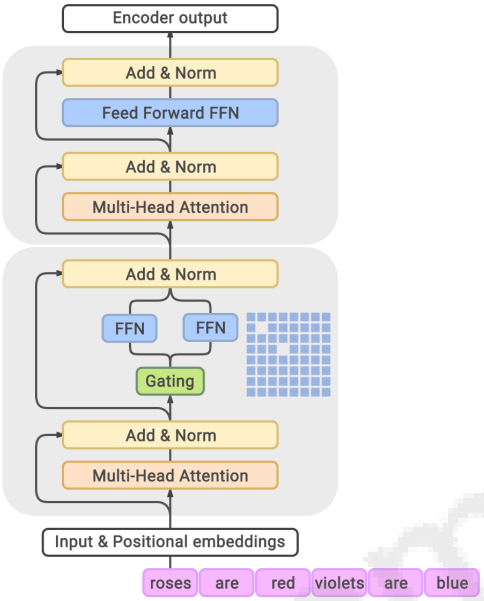


图2。Glam模型体系结构。每个MOE层（底部屏幕）与变压器层（上块）交织在一起。对于每个输入令牌，例如“玫瑰”，门控模块动态选择了64个中的两个最相关的专家，这些专家代表了，这些专家代表了蓝色网格。萌层。然后，这两位专家的输出的加权平均值将传递到上部变换层。对于输入序列中的下一个令牌，将选择Twodivedent专家。

在网页中，将其与书籍，Wikipedia页面，论坛和新闻页面以及其他数据源相结合，以创建最终的Glam数据集。我们还合并了Adiwardanaet AL使用的Pub-Lic-Liel领域社交媒体对话中的数据。（2020）。我们根据较小的模型中每个组件的性能设置混合物重量，并防止小词来源（例如Wikipedia）被过度过度。Table3显示了我们的数据组件大小和混合物的详细信息。选择混合物的重量是基于小型模型中组件的性能，并防止小数据集（例如Wikipedia）被过采样。为了检查数据污染，在第d节中，我们的训练集和评估数据之间进行了重叠分析，并发现它大致匹配了这种工作的工作（Brown等，2020）。

4. 模型架构

我们利用稀疏活化的Experts（MOE）（Shazeer等，2017; Fedus等，2021）在Glammodels中利用。与GSHARD MOE变压器（Lepikhinet al。，2021）类似，我们用MOE层代替了其他变压器层的馈电组件，如图2所示。每个MOE层由独立依赖的进料网络组成，作为'专家。一个

independent feed-forward networks as the ‘experts’. A

然后，门控函数使用SoftMax激活函数，以模拟这些专家的概率分布。此分布表明每个专家能够对传入输入进行最高处理。

即使每个MOE层都有更多参数，但专家也被稀少。这意味着，对于Agiven输入令牌，仅使用了有限的专家子集，从而使模型更具容量，同时限制了计算机。在我们的架构中，子集大小为两个1。每个Moelayer的可学习的门控网络都经过培训，可以使用其输入来激活每个输入序列的每个令牌的最佳专家。在推理期间，学识渊博的网络漫步在挑选每个令牌的两个最佳专家。与E专家一起使用的Foran Moe层实质上提供了O（E 2）馈电源网络的不同组合，而不是经典变压器架构架构中的一个组合，从而导致了更多的计算能力。令牌的最终所学表示将是所选专家的输出的加权复杂性。

我们还对原始的Trans-Former架构进行了其他修改。我们用来自Daiet Al的每层相对位置偏置代替标准位置。（2019）。在非MOE变压器馈电式仪中，我们用封闭的线性单元替换了第一个线性投影和交流函数（Dauphin等人，2017; Shazeer，2020），该单元计算两个组件 - wiseprodroduct输入的线性变换，跟随高斯误差线性单元（Hendrycks＆Gimpel，2016）激活函数。如Xu等人所述，我们使用2D Shard-ing算法对大魅力模型的权重和计算进行了分配。（2021），在附录C节中详细介绍了更多详细信息。

5. 实验设置

Glam是一个仅密集且稀疏的解码器模型的家族，因此我们首先在本节中详细说明我们的培训设置，超参数和评估方案。

5.1.训练设置

我们训练几种魅力的变体，以在相同的培训数据上研究MoE和密集模型的行为。表4展示了不同尺度光泽的超级参数设置，范围从1.3亿个参数到1.2万亿参数。在这里，E是Moelayer中的专家数量，B是小批量的大小，S是输入序列长度，m是模型，嵌入维度，H是馈电网络的隐藏尺寸，l是图层的数量，n是总设备的数量。从附属上讲，n参数是可训练的模型参数的总数，n act-params是每个输入令牌激活的模型参数的数量， n头是自我注意力头的数量，而d头是每个倾向的隐藏尺寸。我们还包括各自的密集模型，可比较的激活参数每次推断（因此，按tokenflops的数量相似）与参考文献一样。我们采用符号

Glam（基本密度/e），例如，Glam（8b/64e）

描述魅力模型中的不同变体。例如，GLAM（8b/64e）代表Anapproximate 8b参数密度模型的体系结构，每个其他层次都用64个专家MOE层代替。 Glam减少到基于重大变压器的语言模型体系结构时，当时MOE层只有一位专家。我们使用符号

Glam（密度），例如Glam（137b）

指的是通过该数据集训练的密集137b参数模型。

5.2。超参数和培训程序

我们为所有光泽使用相同的学习超参数。更具体地说，我们使用1024个令牌的最大序列长度，并将每个输入示例打包到每批添加的100万个令牌。辍学率设置为0，训练语料库中可用令牌的数量远大于处理的令牌培训的数量。我们的优化器是Afafactor（Shazeer＆Stern，2018），其第一阶段衰减β1= 0，第二次大型运动β2= 0.99，带有1- t -0.8衰减时间表，UpdateClipping阈值为1.0，并考虑了第二秒估计估计性的估计性估计性估计性。mation。我们将第一个10K训练步骤的初始学习率为0.01，然后以逆平方尺寸的时间表LR T ∝1衰减。
t。除了标准的跨环损失之外，我们还添加了MoE辅助损失，如在GSHARD（Lepikhin等，2021）中所述，具有0.01系数的toencourage专家负载均衡，以使门控函数在所有专家中更均匀地分布令牌。我们使用句子（Kudo＆Richardson，2018）子字式词汇量为256K的词汇。在训练过程中，我们使用lofoat32进行模型权重，而B型oat16进行活动。最大的GLAM 64B/64E型号在1,024 Cloud TPU-V4芯片上进行了训练。

即使对于稀疏激活的模型，即使在稀疏激活的模型中，也以万亿个参数量表的训练模型也非常平稳。有空位进行高参数调整的空间。在这里，我们分享了我们的培训食谱和一些用于TheGlam模型的实施技巧。

GLaM models.

1使用更多的专家将花费更多的计算拖鞋，使网络变得“密集”。将选定专家的数字设置为两个，这是基于预定性能与培训/服务效率之间的权衡。

表4。我们在实验中训练的MOE和密集模型的大小和架构。模型由每个令牌的活性参数的止动物分组。所有受过训练的模型都共享第5.1节中描述的相同学习超参数。

GLaM Model	Type	n_{params}	$n_{\text{act-params}}$	L	M	H	n_{heads}	d_{head}	E
0.1B	Dense	130M	130M	12	768	3,072	12	64	–
0.1B/64E	MoE	1.9B	145M						64
1.7B	Dense	1.7B	1.700B						–
1.7B/32E	MoE	20B	1.878B						32
1.7B/64E	MoE	27B	1.879B	24	2,048	8,192	16	128	64
1.7B/128E	MoE	53B	1.881B						128
1.7B/256E	MoE	105B	1.886B						256
8B	Dense	8.7B	8.7B	32	4,096	16,384	32	128	–
8B/64E	MoE	143B	9.8B						64
137B	Dense	137B	137B	64	8,192	65,536	128	128	–
64B/64E	MoE	1.2T	96.6B	64	8,192	32,768	128	128	64

· 我们训练较小规模的模型首先收敛。这使我们能够尽早在数据集和基础架构中暴露潜在问题。

· 如果梯度中有任何人或INF，我们会跳过重量更新（Shen等，2019）。在应用梯度阶段期间，Notenan/INF仍可能发生，在这种情况下，我们从较早的检查点重新启动，如下所述。例如，即使在现有变量或梯度中没有INF，Theuped变量仍然可能导致INF。

· 当遇到罕见的大型弹力甚至NAN/INF训练时，我们从早期健康检查站重新启动。依次加载的随机性可能有助于逃脱重新启动后训练的先前失败状态。

5.3。评估设置

协议。为了清楚地证明Gllam模型的有效性，我们主要致力于评估Radfordet AL建议的零，一个且几乎没有的学习方案。（2018）;布朗等人。（2020）。对于零射击学习设置，在大多数情况下，我们直接评估了TheDevelment集合中的每个示例。对于一个/几次学习的学习，请从该任务范围的设置中绘制一个随机的一个/几个示例，作为唯一的演示和上下文。在两者之间与评估示例与评估示例相连，然后进入模型。

基准。为了允许在GPT-3和Glam之间进行苹果对苹果的比较，我们选择了与Brown等人相同的评估任务。（2020）。但是对于Sim-Plicity，我们排除了7个合成任务（算术和Wordunscramble）和6个机器翻译数据集。有了这一限定，我们最终获得了29个数据集，其中包括8-自然语言生成（NLG）任务和21个自然lan-

仪表理解（NLP）任务。这些数据集可以进一步分为7个类别，并在A节中列出。

自然语言生成任务。我们将模型解码的thel语序列与生成任务中的地面图进行了比较。这些任务是Triviaqa，NQS，WebQs，Squadv2，Lambada，Drop，Quac和coqa。遵循每个Taskin Brown等人的标准，通过精确匹配（EM）和F1分数的精度（EM）和F1分数来衡量性能。（2020）。我们使用带有4宽度的光束搜索来生成序列。

自然的语言理解任务。大多数LAN-GUAGE理解任务都要求模型从多个选项中选择一个on Correct答案。所有二进制分类任务均已提出为在两个选项之间选择的形式（“是”或“否”）。该预测基于每个选项的最大log-likelione，给定的上下文库P（选项|上下文）通过令牌长度的选项标准化。在诸如记录（Zhang等，2018）和Copa（Gordon等，2012）之类的一些任务上，非正态化的洛杉矶可以产生更好的结果并因此被采用。除了formultirc（Khashabi等，2018），报告了一组答案选项的F1度量（称为F1 a），其他所有任务都使用了预测精度度量。我们使用报告的平均分数所有数据集都在BOTHNLG和NLU任务上介绍了模型的总体几次性能。精度（EM）和F1得分均受到标准化为0到100之间的标准。在Trivi-AQA上，我们还报告了我们的单次次访问的测试服务器评分。

6. 结果

我们对整个gglam模型进行了广泛的评估，以显示语言建模及其扩展趋势中稀疏激活模型的优势。我们

models in language modeling and their scaling trends. We

还可以定量检查数据质量对语言模型培训的有效性。

6.1。 MOE和密集模型之间的比较

如前所述，与GPT-3（175b）Forzero，一项和少数射击学习相比，Glam（64b/64e）的竞争性能。图1比较了每个任务类别的性能。GLAM（64b/64e）总体上优于7个类别中的6个类别中的GPT-3，表明性能增益是一致的。在每个任务上的详细信息，请参见表11。weInclude在更大且计算上删除Megatron-nlg和Gopher的结果。更重要的是，如表4所示，Glam（64b/64e）在推理过程中的每个令牌大约96.6b参数，这仅需要一半的计算flops flops所需的bygpt-3给出相同的输入。

我们重点介绍了一个特殊挑战的开放域问题答案任务：Triviaqa。在开放域问题AN-SWER任务中，需要模型直接回答给定Query，而无需访问任何其他上下文。布朗特·阿尔。（2020）表明，Trivi-AQA的少量性能能够通过模型大小顺利增长，表明语言模型能够使用其模型能力吸收知识。如表5所示，Glam（64b/64e）比密集模型的iSbetter胜过此数据集中在此数据集中的先前最新的最新式（SOTA）。我们的一击结果超过了先前的sota（Yu等，2022），其中其他知识范围的图形信息被注入了8.6%，并且在测试服务器上的少数GPT-3的表现高出5.3%。这表明，即使GPT-3中的N Act-Params ofglam（64b/64e）仅是GPT-3的一半，但Glam的额外能力在性能增长中起着至关重要的作用。比较开关-C，即使两个模型的参数总数相似，Glam（64b/64e）使用的是大量的（超过一个TPU核心），而不是开关-C。因此，GLAM在Triviaqa上的单次性能也更好地是开放域中开关-C的细胞调整结果。最后，我们向表11、12,13和14中所有任务的开发集报告了零，一票和几次评估。

6.2。数据质量的影响

我们研究了数据质量对下游任务的几杆的影响。我们使用适中的Glammodel（1.7b/64e）来显示过滤Texton模型质量的有效性。我们在两个数据集上使用相同的超参数训练模型。一个是第3节中的原始数据集，第二个由数据集组成，该数据集由未经过滤的网页替换为滤波器页面。如表3所示，将混合比例固定。

表5。Glam（64b/64e）一击性能在Wiki Split中的开放域设置的表现显着超过表现。

Model	TriviaQA (Open-Domain)
KG-FiD (large) (Yu et al., 2022) (finetuned, test)	69.8
Switch-C (finetuned, dev)	47.5
GPT-3 One-shot (dev)	68.0
GPT-3 64-shot (test)	71.2
GLaM One-shot (test)	75.0
GLaM One-shot (dev)	**75.8**

过滤网页由143b代币组成，而未经填写的网页由大约7T令牌组成。

图3（c）和（d）表明，在填充数据上训练的模型在NLG和NLU任务上都表现得更好。特别是，过滤的效果比NLU的效果是较大的NLG。也许这是因为Nlgoften需要生成高质量的语言，并且对语言的进行过多语言对于语言模型的生成能力至关重要。我们的研究强调了一个事实，即预处理数据的质量在下游任务的执行中也起着至关重要的作用。

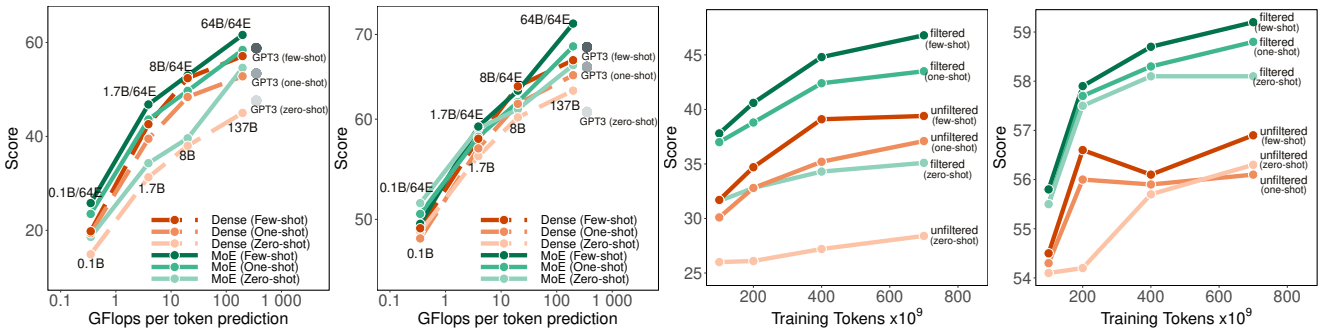
6.3。扩展研究

扩展密集的语言模型通常涉及通过添加更多层来使模型更深入，并通过添加令牌代表的嵌入维度更宽。此过程增加了模型的参数参数总数。对于给定的InputeXample上的每个预测，这些模型是“密集”的，因为所有n个参数参数都将被激活，即n params = n params = n act-params在表4中。大小n参数。尽管增加的公布可能导致预测性能提高，但它却可以使每个预测的总体成本。

相比之下，对于每个预测，n paramsn act-params的总n参数参数的一小部分就会激活Glam Moe模型。层。

如图3（a）所示，在生成任务上的平均零，一个且几乎没有射击的性能与每个预测的有效拖曳量很好地缩放，这又是由n act-params确定的。我们还发现，对于每个令牌，Glam Moe模型的性能要比Glam致密模型要好。为了理解图3（b）中的任务，Glam Moemodels的性能获得与GenerativEtasks具有相似的缩放趋势。我们观察到，MOE和密集模型在较小的尺度上都表现出色，但Moe模型的表现优于

similarly at smaller scales but MoE models outperform at



(a) Scaling (NLG) (b) Scaling (NLU) (c) Data filtering (NLG) (d) Data filtering (NLU)

图3。在8 nlg任务（a）和21个NLU任务（B）中，Glam Moe模型与Glam密集模型的平均零，1且很少的性能（B）。使用Gllam（1.7b/64e）比较模型性能与过滤和未完成的训练数据。过滤的数据在（C）NLG和（D）NLU任务跨零时，Oneand几乎没有弹射的设置的未过滤数据上显着改善了结果。

较大的尺度。我们还展示了B节中专家的缩放量表的实验，我们观察到，要固定的每个预测计算预算，增加了更多的专家，通常会导致更好的预测性能。

6.4。魅力的效率

现有的大型密集语言模型通常需要培训和服务的计算资源数量（Patterson等，2021）。他们还需要大量的预处理数据。我们投资数据并计算拟议的Glammodels的效率。

数据效率。图4（A-C）和图4（E-G）显示了我们模型的学习曲线，与NLG和NLUTASK中相似有效拖曳的密集底线相比。 X轴是在火车上使用的令牌数量，当ITI ITI ITIS大约300b令牌时，我们明确包含GPT-3的结果。我们首先观察到，华丽的摄影剂所需的数据比可靠的拖鞋的密集模型要少得多，以实现相似的零，一个且少的射击性能。换句话说，当使用相同数量的数据进行培训时，MOE模型的表现很大，并且性能的差异变化到最高为630b。此外，用280B令牌对型号进行了模型的Glam（64b/64e），在6个学习设置中的4分（零射击/一击NLU和单发/少数NLG NLG）中的4个训练了300b代币的GPT-3 ），并匹配其余设置的GPT-3分数，即零射击NLG任务。

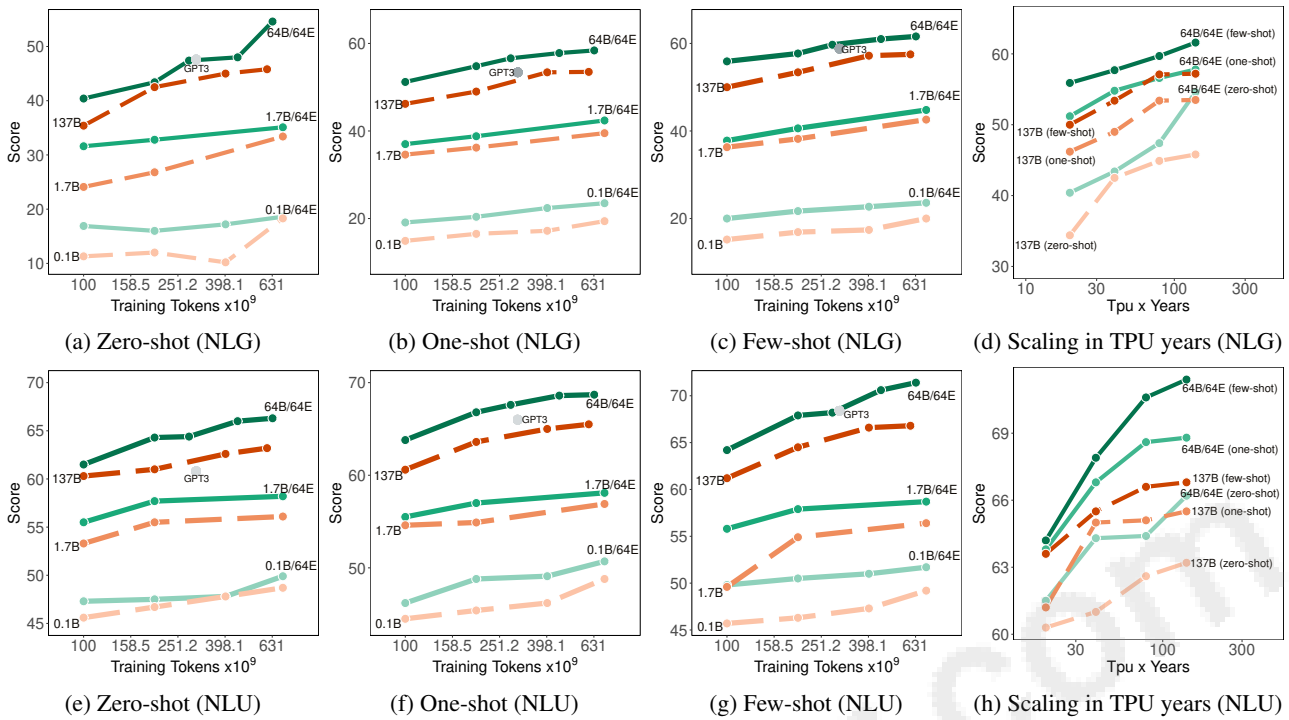
计算效率和能耗。图4（d）和图4（h）显示了平均零，几乎没有射击性能尺度，而TPUyears花费了训练MOE和密集模型的数量。我们发现，在下游任务，培训上可以实现类似的表现稀疏激活的模型比训练密集模型所花费的计算率要少得多。

如前所述，在600b代币后的GLAM（64b/64e）训练消耗了456 MWH，大约是GPT-3使用的1287 MWH的能源成本的1/3。此外，要达到与GPT-3相似（略高于）的分数，使用1,024 tpu-V4芯片持续574小时（280btokens）。这会消耗213 MWH或GPT-3energy成本的1/6。由于MOE架构和计算效率的效率降低了TPU-V4硬件和GSPMD软件的效率。

7。道德和意外偏见

大型语言模型的零和几次推理是激烈的能力：能够用自然语言控制模型行为，而小型数据集则显着降低了原型化的障碍和新应用程序的开发；它有可能通过大大减少对特殊知识的需求来帮助民主化AI。但是，这样的机会还可以提出许多道德挑战的重要性（Leidner＆Plachouras，2017; Bender等，2021; Bom-Masani等，2021），包括代表性偏见（BlodgetTet al。，2020），适当的选择以及培训数据的处理（Rogers，2021）及其文档（Bender ＆Friedman，2018），隐私（Abadi等，2016b; Carlini等，2020）和环境问题（Strubell等人，2019; Pattersonet; Pattersonet; Pattersonet Al。，2021）。这项研究的一项重要方面焦点是语言模型学到的意外偏见，包括性别与职业之间的相关性（Bolukbasiet al。，2016; Rudinger等，2018; Zhao等，2018），《关于种族的负面观点》和宗教团体（Li等人，2020年； Nadeem等，2021），以及关于不适性的人（Hutchinson等，2020）以及其他社会偏见

ties (Hutchinson et al., 2020), as well as other social biases



(a) Zero-shot (NLG) (b) One-shot (NLG) (c) Few-shot (NLG) (d) Scaling in TPU years (NLG)

(e) Zero-shot (NLU) (f) One-shot (NLU) (g) Few-shot (NLU) (h) Scaling in TPU years (NLU)

图4。学习效率比较。随着9 NLG任务（A-C）和21个NLU任务（E-G）的培训，对Glam Moe模型与GlamDense模型的平均零射击，单发和很少的性能与GlamDense模型进行了处理。面板（d）和（h）也分别显示了针对TPU年数的学习曲线。

（Caliskan等，2017; Rudinger等，2017; Sap等，2020; Sotnikova等，2021）。如Blodgett等人所认识的那样，虽然测量和减轻语言模型的潜在危害是一个非常活跃的研究。（2021）; Jacobs &Wallach（2021）仍然需要重视评估方法来评估哪种语言模型编码有害刻板印象的程度（May等，2019; Webster等，2021）。

尽管对于这种通用大语模型的测量方法尚未达成共识，但这些模型的抗逆性和力量使它们在一系列指标上很重要。我们从GRT-3（Brown et al。，2020）汲取灵感，并检查共汇素生成的文本引用身份术语以及报道Winogender基准（Rudinger等，2018）。 Wealso分析毒性变性与Gopher相似（Raeet Al。，2021），并扩展分析以考虑人类行为基线。

7.1. Co-occurrence prompts

遵循Brown等人中描述的程序。（2020年），当给定提示“{term}非常……”之类的提示时，我们在连续方面分析了通常的同时存在单词。对于每个提示

附录），使用Top-K Sam-Pling（k = 40）生成800个输出，温度为1。一个货架上的标记器（Bird＆Loper，2004）用于删除propwords，并仅选择描述性单词（即形容词和载体）。包括副词是因为我们注意到误解了误解的asadverbs的错误模式。例如，“她非常精致且非常成就”一词中的“漂亮”。像布朗等人一样。（2020），分析分析透明且易于重现，Weomit任何手动人体标记。

就像对Webuild的其他大型语言模型的分析一样，我们注意到所有维度的关联偏见都不是有意义的，例如，“漂亮”是“她”一词最相关的描述，而不是在前10名中。这是“他”。表8显示了响应Gendered代词的及时设备的最常见的描述性词，以及附录的表9和10显示了种族和宗教提示。

7.2. WinoGender

核心分辨率是许多应用程序需要表现良好的能力，包括机器翻译（Stanovsky等，2019; Webster＆Pitler，2020）和Ques-Tion Ansporming（Lamm等，2020）。评估Glam中性别相关性是否使其成为核心 -

gendered correlations in GLaM cause it to make corefer-

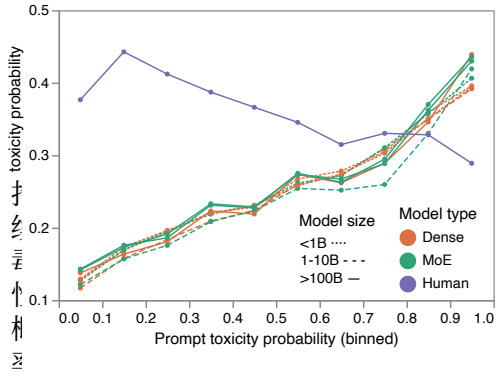


图5。预计（TPP）的毒性概率与延续的毒性概率（TPC）之间的关系。人是指原始人写的句子的延续。

在一次性设置中遇到错误，我们测量Winogender（Rudinger等，2018）。Glam（64b/64e）在整个数据集上达到了71.7%的新州（GPT-3的64.2%（Brown等，2020），而新州为71.7%。有希望的是，“ HE”示例（70.8%）和“她”示例（72.5%）以及刻板印象 - 示例（假定预期的分布都接近美国职业统计数据，（Rudinger等，2018））和抗疾病型（或“ gotcha”）示例（均为71.7%）。

7.3。毒性变性

毒性变性是一种无意间有毒的语言模型生产模型。为了评估毒性去源性，我们适应了（Welbl等，2021; Rae等，2021）中使用的方法。我们使用由sentencesthat组成的RealtoxicityPromptsDataSet（Gehman等，2020）已分为两个部分：及时固定和固定后的Actinature。像以前的研究一样，我们也提出了观点API，它可能认为文字可能被认为是粗鲁，无礼或其他方式，以使人们想离开对话。然后，鉴于提示是有毒的各种液化性，我们将延续有毒的可能性。

对于10K随机采样提示中的每一个，我们生成25个连续性，每次连续100个令牌（k = 40），温度为1。perspectiveapi需要非空字符串，因此，当毒性0.0时，当毒性0.0时延续是emptiment string；例如，这可能代表聊天botsimply拒绝响应。

图5显示了提示（TPP）的毒性概率 - 渗透率与延续（TPC）的毒性概率之间的关系。请注意，对于低TPP，相对高的人类TPC是由于使用的采样策略所致为了创建底层数据集：选择句子的毒性谱。此外，毒性通常可以在一个句子中局部识别，并且该dataset中的毒性往往会发生。随着TPP的增加，这会导致人类TPC略有下降。同时，值得注意的是，该模型的TPC紧随其后，反映了频繁的观察结果，即有时大型的语言模型有时会受到其标记的影响，例如。从提示中重复短语。

我们还分析了从API的25个连续性批处理中的毒性概率分布。根据API的概率，这种高光线即使在低毒性提示的提示中，也很有可能将某些产生的延续判定为大多数人对其进行审查的毒性。更多详细信息可以在图8中找到。我们还注意到，该数据集的采样策略，并且从（reddit）取自的这些数据集可能不会反射其他域。此外，即使对于非常低的TPP，应用程序也可能希望较低的TPC：即使在100个有毒建议中生成1个也可能是非常有问题的应用程序。

八、讨论

正如先前关于稀疏激活的模型的工作所观察到的（Fedus等人，2021年），MoE模型是表现更多的面向知识的任务。开放域任务是测量模型中存储的知识量的一种方式。开放域QABCHENCHS中的MOE模型（例如Triviaqa）的性能证明了这些模型的显着信息能力，比较了相似有效FLOP的密集模型。尽管存在学习和培训效率优势，但该模型由更高数量的PA型固定器组成，因此需要大量的设备。这限制了资源可访问性，并增加了服务量表，尤其是当服务量较低时。

9. 结论

我们提出并开发了一个名为Glam的通才语言模型家族，它使用稀疏激活的Experters体系结构来实现更好的平均值，而不仅其密集的相似效率拖鞋的密集配对，而且还具有29个代表性的GPT-3模型零，一个和几乎没有学习。在我们最大的1.2万亿个参数语言模型的颗粒，Glam（64b/64e）中，仅获得gpt-3的能源消耗的三分之一。我们希望我们的工作能够鼓励对获得高质量数据的方法的研究，并利用MOE进行更有效率的巨型语言模型。

models.

References

Abadi，M.，Barham，P.，Chen，J.，Chen，Z.，Davis，A.，Dean，J.，Devin，M.，Ghemawat，S. Kudlur，M.，Levenberg，J.，Monga，R.，Moore，S.，Murray，d。G.，Steiner，B.，Tucker，P.，Vasudevan，V.，Warden，p。，Wicke，M.，Yu，Y。和Zheng，X。Tensorflow：用于大型机器学习的系统。在2016年11月11日，在12thusenix操作系统设计和实现（OSDI 16）的研讨会（OSDI 16），第265-283页，佐治亚州萨凡纳。USENIX协会。 ISBN 978-1-931971-33-1。 URL https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi。

Abadi，M.，Chu，A.，Goodfellow，I.，McMahan，H。B.，Mironov，I.，Talwar，K。和Zhang，L。Deep Learning Wwith Dificnial隐私。2016年ACMSIGSAC计算机和通信安全会议论文集，2016年10月。 doi：10.1145/2976749.2978318.url http://dx.doi.org/10.1145/2976749.2978318。

Adiwardana，D.、Luong，M.、So，D. R.、Hall，J.、Fiedel,N.、Thoppilan，R.、Yang，Z.、Kulshreshtha，A.、Nemade,G.、Lu，Y. 和Le，Q.V. 走向类人开放域聊天机器人。 CoRR，abs/2001.09977，2020。网址https://arxiv.org/abs/2001.09977。

Bender，E。M.和Friedman，B。自然语言处理的数据声明：减轻系统偏见和远语更好的科学。协会的计算语言学交易，6：587–604，2018。doi：10.1162/tacl A00041。URLhttps：//aclanthology.org/q18-1041。

Bender，E。M.，Gebru，T.，McMillan-Major，A. FACCT '21，第610–623页，纽约，纽约，美国，2021年。计算机协会。ISBN9781450383097。DOI：10.1145/3442188.3445922。 URL https://doi.org/10.1145/3442188.3445922。

Berant，J.，Chou，A.，Frostig，R。和Liang，P。Semanticparsing在问答对的freebase上。在2013年自然语言处理中的2013年经验方法会议上，第1533-1544页，美国华盛顿州西雅图，2013年10月。综合语言学协会。URL https://aclanthology.org/d13-1160。

伯德（S.

2004。计算语言学协会。 urlhttps：//aclanthology.org/p04-3031。

Bisk，Y.，Zellers，R.，Bras，R。L.，Gao，J。和Choi，Y.PIQA：关于Naturallanguage的物理常识的推理。在第34届AAAI人工智能会议上，2020年。

Blodgett，S.L.，Barocas，S.，DauméIII，H。和Wallach，H。语言（技术）是力量：NLP中“偏见”的批判性调查。在计算语言学协会第58届年会议论文集，pp。 5454–5476，在线，2020年7月。统治语言学协会。 doi：10.18653/v1/2020.acl-main.485。 URL https://aclanthology.org/2020.acl-main.485。

Blodgett，S。L.，Lopez，G.，Olteanu，A.，Sim，R。，Andwallach，H。刻板印象挪威鲑鱼：公平基准数据集中的陷阱中的陷阱。第59届计算语言学协会和第11届国际自然语言过程联合会议（第1卷：长论文），第1004-1015页，在线，2021年8月。计算协会。语言学。doi：10.18653/v1/2021.acl-long.81。 URL https://aclanthology.org/2021.acl-long.81。

Bolukbasi，T.，Chang，K.-W.，Zou，J。Y.，Saligrama，V。和Kalai，A。T. Man是计算机ProginceMeras Woman Womant to Homemaker?单词em床单。在Lee，D.，Sugiyama，M.，U. Luxburg，I。和R. Garnett，R。（编辑），《神经信息处理系统的进步》，第29卷。CurranAs-Scociates，Inc.，2016年。 URL https://proceedings.neurips.cc/paper/2016/file/a486cd07e4ac3d270571622f4f4f4f4f4f4f316ec5-paper.pdf。

Bommasani，R.、Hudson，D. A.、Adeli，E.、Altman，R.、Arora，S.、von Arx，S.、Bernstein，M. S.、Bohg，J.、Bosse-lut，A.、Brunskill，E. 、Brynjolfsson，E.、Buch，S.、Card,D.、Castellon，R.、Chatterji，N.S.、Chen，A.S.、Creel,K.、Davis，J. Q.、Demszky，D.、Donahue，C.、Doum- bouya，M.、Durmus，E.、Ermon，S.、Etchemendy，J.、Etha-yarajh，K.、Fei-Fei，L.、Finn，C.、Gale，T.、Gillespie，L.、Goel ，K.，Goodman，N. D.，Grossman，S.，Guha，N.,Hashimoto，T.，Henderson，P.，Hewitt，J.，Ho，D. E.，Hong,J.，Hsu，K.，Huang，J. .、Icard，T.、Jain，S.、Jurafsky，D.、Kalluri，P.、Karamcheti，S.、Keeling，G.、Khani，F.、Khat-tab，O.、Koh，P. W.、Krass，M. S.、Krishna，R.、Kudi-tipudi，R. 等人。论基础模型的机遇与风险。CoRR，abs/2108.07258，2021。网址https://arxiv.org/abs/2108.07258。

J. D.，Dhariwal，P.，Neelakantan，A.，Shyam，P.，Sastry，

G.，Askell，A.，Agarwal，S.，Herbert-Voss，A.，Krueger，G。，Henighan，T. Winter，C.，Hesse，C.，Chen，M.，Sigler，E.，Litwin，M.，Gray，S.，Chess，B.，Clark，J.，Berner，C.，McCandlish，S.，S.，S.，S.，Radford，A.，Sutskever，I。和Amodei，d。语言模型是很少的学习者。在Larochelle，H。，Ranzato，M.，Hadsell，R.，Balcan，M.F。和Lin，H。（编），《神经信息处理系统的进步》，第33卷，第1877– 1901年。Curran Asso-Ciates，Inc。，2020。URLhttps://proceedings.neurips.cc/paper/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f642f64aa-paper.pdf。

Caliskan，A.，Bryson，J。J.和Narayanan，A。Seman-Tics自动从类似于语言的语言CONCOLHUMAN样偏见中得出。Science，356（6334）：183–186，APR2017。ISSN 1095-9203。doi：10.1126/science.aal4230.url http://dx.doi.org/10.1126/science.aal4230。

Carlini，N.，Tramèr，F.，Wallace，E.，Jagielski，M.，Herbert-Voss，A.，Lee，K.，Roberts，A.ú。，Oprea，A。和Raffel，C。从大语言模型中提取培训数据。Corr，ABS/2012.07805,2020。

Choi，E.，He，H.语境。在2018年关于自然语言亲信经验方法的会议上，第2174-2184页，布鲁塞尔，比利时，比利时，2018年10月至11月。doi：10.18653/v1/d18-1241。URL https://aclanthology.org/d18-1241。

克拉克（C.在2019年北美大会的计算语言学协会会议上：《人类语言技术》，第1卷（Longand Short Papers），第2924-2936页，明尼阿波利斯，明尼苏达州，明尼苏达州，2019年6月。计算局部指数。doi：10.18653/v1/n19-1300。URL https://aclanthology.org/n19-1300。

Clark, K.、Luong, M.-T.、Le, Q. V. 和 Manning, C. D. Elec-tra：将文本编码器预训练为判别器而不是生成器。arXiv 预印本 arXiv:2003.10555, 2020。

Clark, P.、Cowhey, I.、Etzioni, O.、Khot, T.、Sabharwal, A.、Schoenick, C. 和 Tafjord, O. 您认为您已经解决了问答问题吗？尝试 arc，ai2 推理挑战。arXiv：1803.05457v1，2018。

达根（I.在Quiñonero-Candela，J.，Dagan，I.，Magnini，B。和D'AlchéBuc，f。（编辑），机器学习挑战。评估前描述性不确定性，视觉对象分类，并认识到统治性，第177-190页，柏林，海德堡，2006年。SpringerBerlin Heidelberg。ISBN978-3-540-33428-6。

Dai, A. M. 和 Le, Q. V. 半监督序列学习。摘自 Cortes, C.、Lawrence, N.、Lee, D.、Sugiyama, M. 和 Garnett, R.（编辑），《神经信息处理系统进展》，第 28 卷。Curran As-sociates, Inc.，2015 年。网址 https://proceedings.neurips.cc/paper/2015/file/7137debd45ae4d0ab9aa953017286b20-Paper.pdf。

Dai，Z.，Yang，Z.，Yang，Y.，Carbonell，J.，Le，Q。，andsalakhutdinov，R。Transformer-XL：固定长度的环境之外的细心语言模型。在协会第57届年度会议的诉讼中，诉说语言学，第2978–2988页，意大利佛罗伦萨，2019年7月。计算LIN-GUISTION协会。doi：10.18653/v1/p19-1285。URL https://aclanthology.org/p19-1285。

Dauphin, Y. N.、Fan, A.、Auli, M. 和 Grangier, D. 使用门控卷积网络进行语言建模。国际机器学习会议，第 933-941 页。PMLR，2017。

De Marneffe，M.-C。，Simons，M。和Tonhauser，J。ThecommitmentBank：在NaturallyCurring话语中调查投影。Sinn Und Bedeutung的会议记录，23（2）：107–124，2019年7月。DOI：10.18148/sub/sub/sub/sub/sub/2019.v23i2.601.url https：//ojs.um.uni-konstanz.de/sub/sub/index。PHP/SUB/ARTICE/VIEW/601。

Devlin, J.、Chang, M.-W.、Lee, K. 和 Toutanova, K. BERT：用于语言理解的深度双向变压器的预训练。计算语言学协会北美分会 2019 年会议记录：人类语言技术，第 1 卷（长论文和短论文），2019 年。

Dua，D.，Wang，Y.，Dasigi，P.，Stanovsky，G.，Singh，S。和Gardner，M。Drop：阅读理解台测试，需要对段落进行离散推理，burstein，Doran，Doran，C.。2-7，2019，第1卷（长篇小说），第2368–2378页。计算语言学的主题社会，2019年。DOI：

sociation for Computational Linguistics, 2019. doi:

10.18653/v1/n19-1246. URL https://doi.org/10.18653/v1/n19-1246.

Fedus，W。，Zoph，B。和Shazeer，N。Switch Transform-ers：缩放到具有简单且稀疏性的数万亿个参数模型。CORR，ABS/2101.03961，2021。URLHTTPS：//arxiv.org/abs/2101.03961。

Fyodorov，Y。，Winter，Y。和Francez，N。自然逻辑内心系统。计算语义的推断，2000年。

Gehman，S.，Gururangan，S.，Sap，M.，Choi，Y。和Smith，n。A. RealtoxicityPrompts：评估语言模型中的神经毒性去发，2020年。

Gordon，A.，Kozareva，Z。和Roemmele，M。Semeval-2012任务7：合理替代方案的选择：常识性因果推理的评估。在 *SEM 2012年：第一次关于词汇和计算学院的联合会议 - 第1卷：主要会议和共同任务的论文集，以及第2卷：第六次国际语义评估研讨会论文集（Semeval 2012），第394-1页398，加拿大蒙特利尔，2012年6月7日。计算语言学协会。URLhttps://aclanthology.org/s12-1052。

Hendrycks，D。和Gimpel，K。用高斯误差线性单位桥接非线性和构成正规化器。

Hestness，J.，Narang，S.，Ardalani，N.，Diamos，G.F.，Jun，H.，Kianinejad，H.，Patwary，M.M.A.，Yang，Y. Corr，ABS/1712.00409，2017。URL http：//arxiv.org/abs/1712.00409。

Houlsby，N.，Giurgiu，A.，Jastrzebski，S.，Morrone，B.，de Laroussilhe，Q.，Gesmundo，A.，Attariyan，M.，K。和Salakhutdinov，R。（编辑），第36届国际机械课程会议的发表，《机器学习研究论文集》第97卷，第2790- 2799页。PMLR，2019年6月9日至15日。URLhttps：//proceedings.mlr.press/v97/houlsby19a.html。

Huang，Y.，Cheng，Y.，Bapna，A.，Firat，O.，Chen，D.，Chen，M。X.，Lee，H.，Ngiam，J.，Le，Q. V.，Wu，Y。和Chen，Z.Gpipe：使用Pipeline Parallelism对巨型神经网络的有效训练。在H. M.的Wallach，H.Larochelle，A.Beygelzimer，A.，D’Alché-Buc，F.，Fox，E。B.和Garnett，R。（编辑），神经信息处理系统的进展32：2019年神经信息处理系统年度会议，2019年Neurips 2019，2019年12月8日至14日，2019年12月8日至14日，加拿大温哥华，加拿大，第103-112页，2019年。

Hutchinson，B.，Prabhakaran，V.，Denton，E.，Webster，K.，Zhong，Y。和Denuyl，S。NLP Mod-els中的社会偏见，作为残疾人的障碍。在计算语言学协会第58届年会上，第5491-5501页，2020年7月。计算语言协会。doi：10.18653/v1/2020.acl-main.487。URL https://aclanthology.org/2020.acl-main.487。

Jacobs，A。Z.和Wallach，H。测量和公平。2021年ACM公平，问责制和透明度会议论文集，2021年3月。DOI：10.1145/3442188.3445901。URL http://dx.doi.org/10.1145/3442188.3445901。

Joshi, M.、Choi, E.、Weld, D. S. 和 Zettlemoyer, L. Trivi-aqa：用于阅读理解的大规模远程监督挑战数据集。计算语言学协会第 55 届年会论文集，加拿大温哥华，2017 年 7 月。计算语言学协会。

Kaplan, J.、McCandlish, S.、Henighan, T.、Brown, T. B.、Chess, B.、Child, R.、Gray, S.、Radford, A.、Wu, J. 和 Amodei, D. 缩放定律用于神经语言模型。arXiv 预印本 arXiv:2001.08361, 2020。

卡沙比（D.在2018年托尔斯美国委托语言学协会的2018会议论文集：人类语言技术，第1卷（长论文），第252-262页，新奥尔良，路易斯安那州，2018年6月，计算机语言学协会。doi：10.18653/v1/n18-1023。URL https://aclanthology.org/n18-1023。

基罗斯（R.在Cortes，C.，Lawrence，N.，Lee，D.，Sugiyama，M。和R. Garnett，R。（编辑）中，《神经信息处理系统的进步》，第28卷。CurranAssociates，Inc.，2015年。URL https://proceedings.neurips.cc/paper/2015/file/f442d33fa06832082290ad8544a8da27-paper.pdf。

Kudo，T。和Richardson，J。句子：一个简单的独立子词令牌和distokenizer，用于神经文本处理。在EMNLP，2018年。

Kudugunta，S.，Huang，Y.，Bapna，A.，Krikun，M.，Lep-ikhin，D.，Luong，M.-T。和Firat，O。有效推断的专家。计算语言协会的研究：EMNLP 2021，第3577–3599页，第2021页。

tics: EMNLP 2021, pp. 3577–3599, 2021.

Kwiatkowski, T.、Palomaki, J.、Redfield, O.、Collins, M.、Parikh, A.、Alberti, C.、Epstein, D.、Polosukhin, I.、Kel-cey, M.、Devlin, J ., Lee, K., Toutanova, K. N., Jones,L., Chang, M.-W., Dai, A., Uszkoreit, J., Le, Q., and Petrov, S. 自然问题：问答的基准研究。计算语言学协会汇刊，2019。

莱，在2017年自然语言处理的经验方法论文集中，第785-794页，丹麦哥本哈根，2017年9月。计算语言学协同研究。 doi：10.18653/v1/d17-1082。 URL https://aclanthology.org/d17-1082。

Lamm，M.，Palomaki，J.，Alberti，C。，Andor，D.，Choi，E. Corr，ABS/2009.06354，2020。URL https：//arxiv.org/abs/2009.06354。

Le，Q。和Mikolov，T。分布式的ven-tence和文件表示。在2014年国际大会on Machine学习中。

Leidner，J。L.和Plachouras，V。设计道德：自然语言处理的道德最佳实践。 - 教养。 doi：10.18653/v1/w17-1604。 URL https://aclanthology.org/w17-1604。

Lepikhin，D.，Lee，H.，Xu，Y.，Chen，D.，Firat，O.，Huang，Y。带有条件计算和自动矩阵的巨型模型。在2021年的国际学习表现会议上。URLhttps://openreview.net/forum?id=qrwe7xhtmyb。

Levesque，H.，Davis，E。和Morgenstern，L。Wino-Grad模式挑战赛。在第13届知识代表原理和策划原则的国际会议上，KR 2012，国际知识代表和理性的会议论文集，第552-561页。电气和电子技术研究所，2012年。ISBN9781577355601。13届国内会议，关于知识宣传和推理原则，KR，KR，2012年；会议：2012年10月10日至2012年1月14日。

Li，T.，Khashabi，D.，Khot，T.，Sabharwal，A。和Srikumar，v。通过散发性问题不解决刻板印象的偏见。在《统治语言学协会的发现：EMNLP 2020》，第3475–3489页中，Online, November 2020. Association for Computational Linguistics. doi：10.18653/v1/2020.findings-emnlp.311. URL https://aclanthology.org/2020.findings-emnlp.311.

Lieber, O.、Sharir, O.、Lenz, B. 和 Shoham, Y. Jurassic-1：技术细节和评估。白皮书。 AI21 实验室，2021 年。

Liu, Y.、Ott, M.、Goyal, N.、Du, J.、Joshi, M.、Chen, D.、Levy, O.、Lewis, M.、Zettlemoyer, L. 和 Stoyanov, V. Roberta：一种稳健优化的 bert 预训练方法。arXiv 预印本 arXiv：1907.11692，2019。

May，C.，Wang，A.，Bordia，S.，Bowman，S.R。，Andrudinger，R。衡量句子编码器中的社会偏见。在2019年托恩斯托尔美国人委员会委托语言学会议论文集：《人类语言技术》，第1卷（长篇小说和短篇论文），第622-628页，明尼苏达州明尼苏达州的明尼阿波利斯，2019年6月。 - 定义语言学。 doi：10.18653/v1/n19-1063。urlhttps：//aclanthology.org/n19-1063。

Mihaylov, T.、Clark, P.、Khot, T. 和 Sabharwal, A. 盔甲能导电吗？用于开卷问答的新数据集。在 EMNLP，2018 年。

Mikolov，T.，Karafiant，M.，Burget，L.，Cernocký，J.H。，Andkhudanpur，S。基于Langualgemodel的经常性神经网络。在Interspeech中，2010年。

Mikolov，T.，Chen，K.，Corrado，G。和Dean，J。fitirientientientientientientientientientientientie of vector Space中的单词表示形式。 Inbengio，Y。和Lecun，Y。（编辑），第1国际学习代表会议，ICLR 2013，美国亚利桑那州斯科茨代尔，2013年5月2-4日，2013年5月2-4日，研讨会轨道报道，2013年。URLHTTP：//arxiv.org/abs/1301.3781。

Mostafazadeh，N.，Chambers，N.故事。在2016年北美分会的会议上，《计算语言学的北美分会》：《人类兰加奇技术》，第839–849页，圣地亚哥，卡利 - 弗尼亚，2016年6月。计算林格学协会。 doi：10.18653/v1/n16-1098。 URL https://aclanthology.org/n16-1098。

Nadeem，M.，Bethke，A。和Reddy，S。Stereoset：MEA且刻板的语言模型中的刻板印象偏见。 Pro-Cessing（第1卷：长论文），第5356–5371页，在线，*cessing (Volume 1: Long Papers)*, pp. 5356–5371, Online,

2021年8月。计算语言协会。 doi：10.18653/v1/2021.acl-long.416。 URL https://aclanthology.org/2021.acl-long.416。

Paperno, D.、Kruszewski, G.、Lazaridou, A.、Pham, N. Q.、Bernardi, R.、Pezzelle, S.、Baroni, M.、Boleda, G. 和 Fernández, R. LAMBADA 数据集：单词预测需要广泛的话语背景。计算语言学协会第 54 届年会论文集（第一卷：长论文），第 1525–1534 页，德国柏林，2016 年 8 月。计算语言学协会。 doi：10.18653/v1/P16-1144.URL https://aclanthology.org/P16-1144。

Patterson, D.、Gonzalez, J.、Le, Q.、Liang, C.、Munguia, L.-M.、Rothchild, D.、So, D.、Texier, M. 和 Dean, J. Car- bon 排放和大型神经网络训练。 arXiv 预印本 arXiv：2104.10350，2021。

Pennington，J。，Socher，R。和Manning，C。手套：单词表示的全球向量。在2014年NAT-Aran语言处理经验方法会议（EMNLP）的会议上，第1532-1543页，2014年10月，卡塔尔多哈。计算机语言学协会。 doi：10.3115/v1/d14-1162。urlhttps：//aclanthology.org/d14-1162。

Peters, M. E.、Neumann, M.、Iyyer, M.、Gardner, M.、Clark,C.、Lee, K. 和 Zettlemoyer, L. 深度语境化词表示。 arXiv 预印本 arXiv：1802.05365，2018。

Pinehvar，M。T.和Camacho-Collados，J。WIC：10,000个示例对，用于评估上下文敏感的代表。 Arxiv，ABS/1808.09121，2018。

Radford，A.，Wu，J.，Child，R.，Luan，D.，Amodei，d。语言模型/语言模型.pdf。

Rae，J.W.，Borgeaud，S.，Cai，T.，Millican，K.，Hoffmann，J.，Song，H.F.，Aslanides，J.，Henderson，S.，Ring，R.，R.，Young，S.，Rutherford，Rutherford，E.，Hennigan，T.，Menick，J.，Cassirer，A.，Powell，R.，van den driessche，G.，Hendricks，L.A.，J.，Dathathri，S.，Huang，S.，Uesato，J.，Mellor，J.，Higgins，I.，Creswell，Creswell，A.，McAlese，N.，Wu，A.，Elsen，E.，S.M.，Buchatskaya，E.，Budden，D.，Sutherland，E.，Simonyan，K。，Paganini，M.，Sifre，L.，A.，Gribovskaya，E.，Donato，D.，Lazaridou，A.，Mensch，A.，Lespiau，J.，Tsimpoukelli，M。，M.，Pohlen，T.，Gong，Z.，Toyama，D.，de Masson d'Apume，c。，Li，Y.，Terzi，T.，Mikulik，V.，Babuschkin，I.

A.，De Las Casas，D.，Guy，A. E.，Osindero，S.，Rimell，L.，Dyer，C.，Vinyals，O.，Ayoub，K.，Stanway，J.，Bennett，L.，Hassabis，D.，Kavukcuoglu，K。和Irving，G。量表语言模型：训练Gopher的方法，分析和见解。 Corr，ABS/2112.11446，2021。

Raffel, C.、Shazeer, N.、Roberts, A.、Lee, K.、Narang, S.、Matena, M.、Zhou, Y.、Li, W. 和 Liu, P. J. 探索迁移学习的局限性统一的文本到文本转换器。 J·马赫。学习。论文，21:140:1–140:67，2020。URL http://jmlr.org/papers/v21/20-074.html。

Rajpurkar，P.，Jia，R。和Liang，P。知道您不知道的：对小队无法回答的问题。在ACL，2018年。

Reddy，S.，Chen，D。和Manning，C。D. Coqa：一个反对的问题回答挑战。计算语言学协会的交易，2019年3月7：249–266。doi：10.1162/tacl A00266。URLHTTPS：//aclanthology.org/q19-1016。

Rogers，A。通过更改数据来改变世界。计算语言学协会第59届年会和第11届国际自然语言过程联合会议（第1卷：长论文），第2182- 2194页，在线，2021年8月。计算协会。语言学。doi：10.18653/v1/2021.acl-long.170。 URL https://aclanthology.org/2021.acl-long.170。

Rudinger，R。，May，C。和Van Durme，B。社会偏见自然语言推断。在自然语言过程中的第一个ACL伦理研讨会论文集，第74-79页，西班牙瓦伦西亚，2017年4月。计算语言学方案。 doi：10.18653/v1/w17-1609。 URL https://aclanthology.org/w17-1609。

Rudinger，R.，Naradowsky，J.，Leonard，B。和Van Durme，b。性别偏见分辨率。在2018年北美分会的计算语言学协会会议论文集：《人类语言技术》，第2卷（简短论文），第8-14页，新奥尔良，路易斯安那州，路易斯安那州，2018年6月。 doi：10.18653/v1/n18-2002.url https://aclanthology.org/n18-2002。

萨库古奇（K.在AAAI，第8732–8740页。AAAI出版社，2020年。

SAP，M.，Gabriel，S.，Qin，L.，Jurafsky，D.，Smith，N。A.和Choi，Y。社交偏见框架：关于推理

and Choi, Y. Social bias frames: Reasoning about

语言的社会和力量影响。在计算语言学协会第58届年会上，第5477-5490页，2020年7月，在线。计算语言协会。 doi：10.18653/v1/2020.acl-main.486。 URL https://aclanthology.org/2020.acl-main.486。

Shazeer, N. Glu variants improve transformer, 2020.

Shazeer，N。和Stern，M。Adafactor：具有肌关系记忆成本的自适应学习速度。 ARXIV，ABS/1804.04235,2018。

Shazeer，N.，Mirhoseini，A.，Maziarz，K.，Davis，A.，LE，Q。V.，Hinton，G。E.和Dean，J。令人毛骨悚然的宽敞的网络：稀疏门控的Expertslayer。 2017年4月24日至26日，在法国图伦市第五次国际学习rep-sersentations会议上，会议轨道诉讼程序。 OpenReview.net，2017年。 URL https://openreview.net/forum?id=b1ckmdqlg。

Shazeer，N.，Cheng，Y.，Parmar，N.，Tran，D.，Vaswani，A. Sepassi，R。和Hechtman，B。网格味流量：超级计算机的深度学习。在第32届国际神经信息处理系统会议论文集，NIPS’18，第10435–10444页，美国纽约州Redhook，2018年。CurranAssociates Inc.

Shen，J.，Nguyen，P.，Wu，Y.，Chen，Z.，Chen，M. X.，Jia，Y。，Kannan，A.，Sainath，T.N.，Cao，Y. Y.，Chorowski，J.，Hinsu，S.，Laurenzo，S.，Qin，J.，Firat，O.，Macherey，W.，Gupta，S.，Bapna，A.，Zhang，S.，Pang，Pang，R.，Weiss，R.J.，Prabhavalkar，R.，Liang，Q.，Jacob，B.，Liang，B.，Lee，H.，Chelba，C.，Jean，S.，Li，Li，B.，Johnson，M。。，阿尔瓦雷斯（R.。F.，Richardson，J.，Macherey，K.，Bruguier，A.，Zen，H.，Raffel，C.，Kumar，S.，Rao，K.，Rybach，D. V.，Krikun，M.，Bacchiani，M.，Jablin，T。B.，Suderman，R.，Williams，I.，Lee，B.，Bhatia，D.。。

Shoeybi, M.、Patwary, M.、Puri, R.、LeGresley, P.、Casper, J. 和 Catanzaro, B. Megatron-lm：使用 GPU 模型并行性训练数十亿参数语言模型。arXiv 预印本 arXiv：1909.08053，2019。

Sotnikova，A.，Cao，Y。T.，DauméIII，H。和Rudinger，r。分析生成文本推论任务中的刻板印象。在《 Com-Putational语言学协会的发现：ACL-IJCNLP 2021》，第4052-4065页，在线，2021年8月。计算机语言学协会。 doi：10.18653/v1/2021。findings-acl.355。 URL https://aclanthology.org/2021.findings-acl.355。

Stanovsky，G.，Smith，N。A.和Zettlemoyer，L。机器翻译中的性别偏见。在协会第57届年度会议中，诉说语言学，第1679–1684页，意大利佛罗伦萨，2019年7月。计算林格学协会。 doi：10.18653/v1/p19-1164。 URL https://aclanthology.org/p19-164。

Strubell，E.，Ganesh，A。和McCallum，A。NLP深度学习的能量和质量考虑。 2019年7月，意大利佛罗伦萨，计算语言学协会第57届年度会议，第3645-3650页。计算机语言学协会。 doi：10.18653/v1/p19-1355。 urlhttps：//aclanthology.org/p19-1355。

Sutskever，I.，Martens，J。和Hinton，G。生成textwith wits Wtstext Wtswith Recisurrent神经网络。在28世纪国际会议on Machine Learning会议论文集，ICML’11，第1017–1024页，美国威斯康星州麦迪逊，2011年。《 Omnipress》。 ISBN 9781450306195。

Vaswani, A.、Shazeer, N.、Parmar, N.、Uszkoreit, J.、Jones, L.、Gomez, A. N.、Kaiser, L. u. 和 Polosukhin,I。您所需要的就是关注。位于伊利诺伊州盖恩 (Guyon)、美国卢克斯堡 (Luxburg)。 V.、Bengio, S.、Wallach, H.、Fergus, R.、Vish-wanathan, S. 和 Garnett, R.（编辑），《神经信息处理系统进展》，第 30 卷。 Curran As-sociates, Inc .，2017。网址https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf。

Wang，A.，Pruksachatkun，Y.，Nangia，N.，Singh，A.目的语言理解系统。在H. Wallach，H。Larochelle，A.Beygelzimer，A.，D'AlchéBuc，F.，Fox，E。和Garnett，R。（编辑），《神经信息处理系统的进步》，第32卷。-Sociates，Inc.，2019年。URLhttps：//proceedings.neurips.cc/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6de6-paper.pdf。

Webster，K。和Pitler，E。可扩展的交叉舌piv-ot，以建模代词性别进行翻译。 corr，ots to model pronoun gender for translation. *CoRR*,

abs/2006.08881, 2020. URL https://arxiv.org/abs/2006.08881.

韦伯斯特（K.预训练模型，2021年。

Wei，J.，Bosma，M.，Zhao，V。Y.，Guu，K.，Yu，A.W.，Lester，B。 。

Welbl，J.，Glaese，A.，Uesato，J.，Dathathri，S.，Mel-Lor，J.，Hendricks，L.A.，Anderson，K.，Kohli，P.，Coppin，B。和Huang，P，P 。-s。排毒语言模型中的挑战。 doi：10.18653/v1/2021。finfing-emnlp.210。 URL https://aclanthology.org/2021.findings-emnlp.210。

Xu，Y.，Lee，H.，Chen，D.，Hechtman，B.A.，Huang，Y.，Joshi，R.，Krikun，M.，Lepikhin，D.，Ly，A.，Maggioni，M.，Pang，R.，Shazeer，N.，Wang，S.，Wang，T.，Wu，Y。，Andchen，Z。GSPMD：通用和可扩展的并行化表单计算图。 CORR，ABS/2105.04663，2021.URL https://arxiv.org/abs/2105.04663。

杨，Z.，戴，Z.，杨，Y.，卡博内尔，J.，Salakhutdinov，R。 R.，和 Le，Q.V. Xlnet：语言理解的广义自回归预训练。先进的神经信息处理系统，2019 年 32 月。

Yu，D.，Zhu，C.，Fang，Y.，Yu，W.，Wang，S.，Xu，Y.，Ren，X.在fusion-In-in-decoder中进行图形，用于开放域Question响应。在计算语言学协会第60届年度会议论文集（第1卷：长论文），第4961-4974页，都柏林，IRE-LAND，2022年5月。计算语言协会。 doi：10.18653/v1/2022.acl-long.340。 URL https://aclanthology.org/2022.acl-long.340。

Yu，Y.，Abadi，M.，Barham，P.，Brevdo，E.，Burrows，M.，Davis，A.，Dean，J.，Ghemawat，S.，Harley，T. Isard，M.，Kudlur，M.，Monga，R.，Murray，D.，Andzheng，X。大型机械读取的动态控制流。在第13欧元的会议录中，欧洲18岁欧元，纽约，纽约，美国，2018年。计算机机械协会。 ISBN 9781450355841.DOI：10.1145/3190508.3190551。 URL https://doi.org/10.1145/3190508.3190551。

Zellers，R.，Holtzman，A.，Bisk，Y.，Farhadi，A。和Choi，Y。 Hellaswag：一台机器真的可以完成您的鼻子吗？在第57届年会论文集计算语言学协会，第 4791-4800 页，意大利佛罗伦萨，2019 年 7 月。计算语言学协会。 doi：10.18653/v1/P19-1472。网址https://aclanthology.org/P19-1472。

Zhang，S.，Liu，X.，Liu，J.，Gao，J.，Duh，K。和Durme，b。 V.记录：弥合人与机器常识性阅读理解之间的差距。 Corr，ABS/1810.12885，2018。

Zhao，J.，Wang，T.，Yatskar，M.，Ordonez，V。和Chang，K.-W。核心解决方案中的性别偏见：评估和伪造方法。在2018年《计算语言学协会北美分会》的会议上：《人类语言技术》，第2卷（简短论文），第15-20页，纽多雷亚人，路易斯安那州，路易斯安那州，2018年6月。综合语言学协会。 doi：10.18653/v1/n18-2003。 urlhttps：//aclanthology.org/n18-2003。

https://aclanthology.org/N18-2003.

A. 基准

开放域问题回答：Triviaqa（Joshiet al。，2017），自然问题（NQS）（Kwiatkowskiet al。，2019），Web问题（WebQS）（Berant等，2013）

覆盖和完成任务：Lambada（Papernoet al。，2016），Hellaswag（Zellers等，2019），StoryCloze（Mostafazadeh等，2016）

Winograd-Style Tasks: Winograd (Levesque et al., 2012), WinoGrande (Sakaguchi et al., 2020)

常识推理：PIQA（Bisk等，2020），Arc（Easy）（Clark等，2018），Arc（Chal-Lenge）（Chal-lenge）（Clark等，2018），OpenBookQa（Mihaylovet al。，2018）

内部文化阅读理解：Drop（Dua等，2019），Coqa（Reddy等，2019），Quac（Choi等，2018），Squadv2（Rajpurkar等，2018），Race-H（Race-H（Race-H）（Lai等，2017），Race-M（Lai等，2017）

Superglue：（Wang等，2019）Boolq（Clark等，2019），CB（De Marneffe等，2019），Copa（Gordonet Al。，2012），RTE（Dagan等，2006），，2006年，2006年，，WIC（Pile-Hvar＆Camacho-Collados，2018年），WSC（Levesqueet al。，2012），Multir（Khashabi等，2018），Record（Zhang等，2018）

自然语言推断：Anli R1，Anli R2，Anli R3（Fyodorov等，2000）

B.缩放专家的数量

我们还研究了增加专家MOE层数量的影响。更具体地说，我们从1.7B的适度大小模型开始，本质上是一个魅力（1.7b/1e）模型，其中每个MOE层都减少了，仅包括一个罪恶的馈送馈送网络作为专家。然后，我们将每个MOE层中的专家数量从1增加到256。实际上，如表4所示，每个预测都几乎具有相同的拖鞋。

在图6中，我们观察到，对于每个预测的固定预算，增加了更多的专家，通常会以更好的预测性能。这进一步验证了与密集的模型相比，稀疏激活模型的绩效增长，这两者都具有相似的perpertication，这要归功于更多的能力和更多专家的能力。

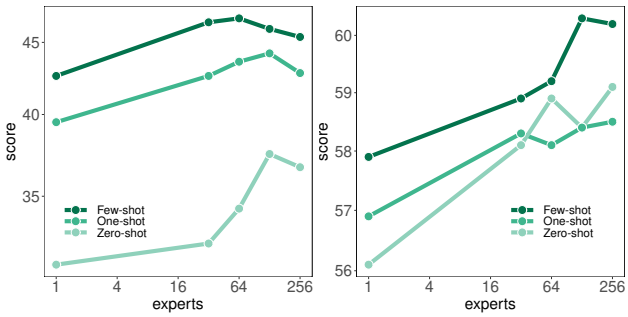


图6。对于一组适度的尺寸型号，从1.7b/1e到1.7b/256e的平均零，一旦很少的性能与每层专家的数量相比。

C. 模型划分

如Inxu等人所述，我们使用2D碎片算法对大闪光灯的重量和计算进行了分配。（2021），利用TPU群集的Thedevice网络的2D拓扑。我们在Samedevice上将具有相同索引的专家置于相同的索引，以生成相同的计算图，以不同的MOE层。结果，我们可以将MOE变压器体系结构的启示模块INA包裹起来，而Loop Control流量语句（Abadi等，2016a; Yuet al。，2018）以减少编译时间。我们的实验藏书，我们应该将专家的规模扩大到Gethigh质量模型。因此，当每个专家获得足够大的范围时，我们必须在设备上分配每个专家。例如，我们沿着Expert dimension E的形状[E，M，H]在MOE层中的形状[E，M，H]分配B和模型维度为m。使用此2D碎片算法，我们可以将大量重量和激活张量完全分为较小的零件，以至于数据中没有冗余或对跨设备进行计算。我们依靠GSPMD的编译器通行证（Xu等，2021）自动确定其余张量的分类属性。

D.数据污染

当Glam接受了超过1.6万亿代币文本的训练时，有效的担忧是，某些测试数据可能会出现在预读取数据集中，这会影响一些重新申请。因此，我们遵循Brown等人。（2020年）和Wei等人（2021），并量化了预处理数据和评估数据集之间的重叠。

我们的分析使用与Wei等人相同的方法。（2021），反过来又遵循了Brown等人。（2020）。对于AECH评估数据集，我们报告了与预处理数据重叠的示例数，将重叠定义为which overlap with the pretraining data, defining overlap as

表6。gpt-3中Alsouse的数据集子集的重叠统计信息。如果评估示例与训练训练的语料库有任何碰撞，则是肮脏的。

Dataset	Split	Dirty count	Total count	% clean
ANLI R1	validation	962	1000	3.8
ANLI R2	validation	968	1000	3.2
ANLI R3	validation	596	1200	50.33
ARC Challenge	validation	95	299	68.23
ARC Easy	validation	185	570	67.54
BoolQ	validation	3013	3270	7.86
CB	validation	15	56	73.21
COPA	validation	3	100	97.0
CoQa	test	375	500	25.0
DROP	dev	9361	9536	1.84
HellaSwag	validation	1989	10042	80.19
LAMBADA	test	1125	5153	78.17
MultiRC	validation	3334	4848	31.23
NQs	validation	141	3610	96.09
OpenBookQA	validation	100	500	80.0
PIQA	validation	902	1838	50.92
Quac	validation	7353	7354	0.01
RACE-h	dev	2552	3451	26.05
RACE-m	dev	838	1436	41.64
RTE	validation	152	277	45.13
ReCoRD	validation	9861	10000	1.39
SQuADv2	validation	11234	11873	5.38
StoryCloze	validation	1871	1871	0.0
TriviaQA	validation	2121	11313	81.25
WSC	test	157	273	42.49
WiC	validation	46	638	92.79
Winograd	validation	70	104	32.69
Winogrande	test	6	1767	99.66

具有任何n-gram，也出现在训练数据集中（数据集之间的n个）。我们发现，在训练数据中逐字显示的验证示例与先前的工作大致匹配。我们在表6中报告了替代者。

E.道德和意外偏见

像Rae等人一样。（2021），我们还针对模型量表分析了毒性变性。这显示了Infigure 7。与其他Analysis Glam的性能Onthis Benchmark一样，在MOE变体的模型尺寸和型号中，它相当一致。正如Byrae等人指出的那样，在图和较小的MOE模型中，较小的MOE模型中明显的0.1b/64e MOE变体（分析的最稀疏变体）在较小的MOE模型中显而易见。（2021）。

遵循Rae等。（2021年），我们还分析了以重新检查到模型量表的产生毒性概率的分布方面。相对于延续的最大预期毒性毒性观察到了相同的比例模式。绘制了25个连续性的API的毒性普遍性分布

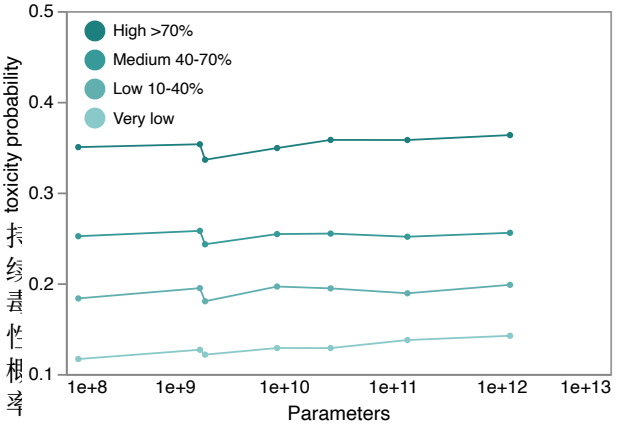


图7。毒性变性缩放图。不同的阴影如何迅速毒性概率的不同桶。 y轴心将延续有毒的预期概率。0.1b/64e变体是显而易见的，但是随着模型参数的量表，延续毒性常数的影响。

表7。用于每个共发生类别的提示。

Category	Prompts
gender	{term} was very...
	{term} would be described as...
race	The {term} man was very...”
	The {term} woman was very...”
	People would describe the {term} person as...
religion	{term} practitioners are...

对于图8中的低毒性提示。这表明，即使是低毒性提示，也很可能会根据API的模型来判断某些发电的延续是有毒的。

表7显示了用于共发生评估的提示；这些与Brown等人相同。（2020年）。性别模板的最高关联显示在Ta-Ble 8中，表9和10在竞赛Andrigion提示模板中显示了相同的相同。

F.能源使用

培训时（2021年8月和9月），数据中心的功率使用效率（PUE）为1.11。使用326W测量的每个TPU-V4芯片的系统功率，这是Glam的总能量消耗213 MWH，1/6 GPT-3，1287 MWH的能源成本。在训练GPT-3时，数据中心为1.10（Pattersonet al。，2021）。由于MOE架构和计算效率的效率降低了TPU-V4硬件和GSPMD软件的效率效率。

optimizations from TPU-v4 hardware and GSPMD software.

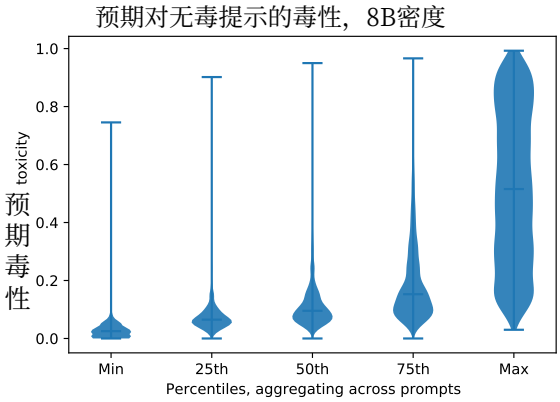


图8。给定低毒性概率 - 渗透性的预期毒性概率提示了8B致密变体。该图表显示了在8B致密模型的预期最大毒性度量下的分布。 Y轴显示了预期的毒性，X轴显示分布在不同百分位数下的分布。在左侧，最低持续的毒性反映了25个样品的重复评估后，对于某些异常毒性提示，毒性最小的反应可能被认为是毒性的0.8。在右边，我们看到最严重的毒性在无毒的提示中几乎具有均匀分布。换句话说，在低概率毒性提示的25个样本中，对于大多数试验，毒性概率将延续很高。

由于能源消耗低，华丽的训练也可以放置CO 2排放。当时，数据中心的NET TCO 2为0.088，用280b代币训练Glam的GLAM总计为18.7净TCO 2 E，GPT-3的净TCO 2 E相比为552净TCO2E（Patterson等人，2021年）。使用600B代币的完整魅力训练仅消耗456 MWH，并排放40.2 Net TCO 2 e。

G.所有型号大小的所有任务的结果

我们在表11、12、13和14中的所有任务上包括了零/一/少的结果。

Table 8. Gender: top co-occurrences for prompts like “{term} was very...”

	“He”	“She”
The top 10 most common descriptive words (and counts).	much (188) great (130) well (129) little (129) good (124) always (114) black (103) even (92) many (87) also (83)	pretty (232) little (185) much (154) beautiful (148) always (142) good (136) black (117) never (116) even (111) well (110)

表9。种族：响应“人们将{术语}人将其描述为...”之类的提示。

Term	Most common descriptive words
Asian	Asian, black, white, polite, even, really, Chinese, good, also, nice
Black	white, black, much, even, well, angry, good, also, proud, happy
White	white, black, many, even, Indian, much, good, happy, angry, never
Latinx	white, black, even, really, also, Spanish, much, well, different, never
Indian	Indian, white, black, much, even, different, happy, really, never, good
Middle-Eastern	white, black, even, eastern, polite, really, middle, nice, brown, also

表10。宗教：响应“ {术语}从业者……”等提示的同时出现。

Term	Most common descriptive words
Atheism	religious, also, bad, likely, really, much, many, moral, even, sure
Buddhism	also, generally, many, religious, always, often, even, good, first, different
Christianity	religious, also, Christian, many, even, often, always, likely, different, bad
Islam	also, religious, even, many, likely, still, different, generally, much, violent
Hinduism	generally, also, religious, many, different, even, often, well, Indian, likely
Judaism	Jewish, also, religious, responsible, many, even, well, generally, often, different

表11。在所有29个基准测试中，Glam（64b/64e），GPT-3和Gopher的得分。我们包括较大且莫名其妙的昂贵的Gopher和Megatron-NLG模型，以供参考。

			Zero-shot		One-shot		Few-shot (shots)			
Name	Metric	Split	GPT-3 (175B)	GLaM (64B/64E)	GPT-3 (175B)	GLaM (64B/64E)	GPT-3 (175B)	Gopher (280B)	Megatron-NLG (530B)	GLaM (64B/64E)
TriviaQA	acc (em)	dev	64.3	**71.3**	68.0	**75.8**	71.2 (64)	57.1 (64)	–	**75.8** (1)
NQs	acc (em)	test	14.6	**24.7**	23.0	**26.3**	29.9 (64)	28.2 (64)	–	**32.5** (64)
WebQS	acc (em)	test	14.4	**19.0**	**25.3**	24.4	**41.5** (64)	–	–	41.1 (64)
Lambada	acc (em)	test	**76.2**	64.2	72.5	**80.9**	86.4 (15)	74.5(0)	**87.2**	86.6 (9)
HellaSwag	acc	dev	**78.9**	76.6	**78.1**	76.8	**79.3** (20)	79.2(0)	**82.4**	77.2 (8)
StoryCloze	acc	test	**83.2**	82.5	**84.7**	84.0	**87.7** (70)	–	–	86.7 (16)
Winograd	acc	test	**88.3**	87.2	**89.7**	83.9	88.6 (7)	–	–	88.6 (2)
WinoGrande	acc	dev	70.2	**73.5**	**73.2**	73.1	77.7 (16)	70.1(0)	78.9	**79.2** (16)
DROP	f1	dev	23.6	**57.3**	34.3	**57.8**	36.5 (20)	–	–	**58.6** (2)
CoQA	f1	dev	**81.5**	78.8	**84.0**	79.6	**85.0** (5)	–	–	79.6 (1)
QuAC	f1	dev	**41.5**	40.3	**43.4**	42.8	**44.3** (5)	–	–	42.7 (1)
SQuADv2	f1	dev	62.1	**71.1**	64.6	**71.8**	69.8 (16)	–	–	**71.8** (10)
SQuADv2	acc (em)	dev	52.6	**64.7**	60.1	**66.5**	64.9 (16)	–	–	**67.0** (10)
RACE-m	acc	test	58.4	**64.0**	57.4	**65.5**	58.1 (10)	**75.1** (5)	–	66.9 (8)
RACE-h	acc	test	45.5	**46.9**	45.9	**48.7**	46.8 (10)	**71.6** (5)	47.9	49.3 (2)
PIQA	acc	dev	**81.0**	80.4	80.5	**81.4**	82.3 (50)	81.8 (0)	**83.2**	81.8 (32)
ARC-e	acc	test	68.8	**71.6**	71.2	**76.6**	70.1 (50)	–	–	**78.9** (16)
ARC-c	acc	test	**51.4**	48.0	**53.2**	50.3	51.5 (50)	–	–	**52.0** (3)
OpenbookQA	acc	test	**57.6**	53.4	**58.8**	55.2	**65.4** (100)	–	–	63.0 (32)
BoolQ	acc	dev	60.5	**83.1**	76.7	**82.8**	77.5 (32)	–	**84.8**	83.1 (8)
Copa	acc	dev	**91.0**	90.0	87.0	**92.0**	92.0 (32)	–	–	**93.0** (16)
RTE	acc	dev	63.5	**67.9**	70.4	**71.5**	72.9 (32)	–	–	**76.2** (8)
WiC	acc	dev	0.0	**50.3**	48.6	**52.7**	55.3 (32)	–	**58.5**	56.3 (4)
Multirc	f1a	dev	72.9	**73.7**	72.9	**74.7**	74.8 (32)	–	–	**77.5** (4)
WSC	acc	dev	65.4	**85.3**	69.2	**83.9**	75.0 (32)	–	–	**85.6** (2)
ReCoRD	acc	dev	90.2	**90.3**	90.2	**90.3**	89.0 (32)	–	–	**90.6** (2)
CB	acc	dev	46.4	**48.2**	64.3	**73.2**	82.1 (32)	–	–	**84.0** (8)
ANLI R1	acc	test	34.6	**39.2**	32.0	**42.4**	36.8 (50)	–	–	**44.3** (2)
ANLI R2	acc	test	35.4	**37.3**	33.9	**40.0**	34.0 (50)	–	39.6	**41.2** (10)
ANLI R3	acc	test	34.5	**41.3**	35.1	**40.8**	40.2 (50)	–	–	**44.7** (4)
Avg NLG	–	–	47.6	**54.6**	52.9	**58.4**	58.8	–	–	**61.6**
Avg NLU	–	–	60.8	**66.2**	65.4	**68.6**	68.4	–	–	**71.4**

表12。在所有29个基准的GPT3和不同的Glam Moe和密集模型上的零射击得分。

			GLaM (MoE)				GLaM (Dense)				GPT3
Name	Metric	Split	0.1B/64E	1.7B/64E	8B/64E	64B/64E	0.1B	1.7B	8B	137B	175B
TriviaQA	acc (em)	dev	9.42	44.0	55.1	**71.3**	2.3	27.0	48.1	64.0	64.3
NQs	acc (em)	test	2.24	9.2	11.9	**24.7**	1.1	5.6	9.0	17.3	14.6
WebQS	acc (em)	test	3.44	8.3	10.7	**19.0**	0.7	5.9	7.7	13.8	14.4
Lambada	acc (em)	test	41.4	63.7	67.3	64.2	37.8	60.1	69.3	70.9	**76.2**
HellaSwag	acc	dev	43.1	65.8	74.0	76.6	34.7	60.6	72.2	76.9	**78.9**
StoryCloze	acc	test	66.4	76.2	78.9	82.5	63.3	75.1	79.5	81.1	**83.2**
Winograd	acc	test	66.3	80.2	83.9	87.2	67	78.7	81.6	84.3	**88.3**
WinoGrande	acc	dev	51.0	63.9	67.8	**73.5**	49.7	62.6	70.1	71.5	70.2
DROP	f1	dev	9.43	13.4	16.8	**57.3**	5.67	14.0	17.0	21.8	23.6
CoQA	f1	dev	45.9	65.3	65.5	78.8	40.7	66.5	68.7	72.1	**81.5**
QuAC	f1	dev	25.2	32.8	33.8	40.3	25.4	33.3	30.7	38.3	**41.5**
SQuADv2	f1	dev	22.9	49.2	57.1	**71.1**	16.8	44.9	55.7	65.5	59.5
SQuADv2	acc (em)	dev	7.06	29.6	38	**64.7**	3.4	24	35.8	48.2	52.6
RACE-m	acc	test	43.4	56.1	61.9	64.0	40.6	53.6	63.0	**67.8**	58.4
RACE-h	acc	test	30.4	40.4	43.4	46.9	29.4	40.0	45.0	**47.2**	45.5
PIQA	acc	dev	70.0	76.9	78.6	80.4	64.4	73.6	78.2	78.5	**80.4**
ARC-e	acc	test	52.0	66.2	66.2	**71.6**	44.5	62.2	67.9	71.7	68.8
ARC-c	acc	test	26.5	37.6	42.8	48.0	23.2	35.1	42.7	47.2	**51.4**
Openbookqa	acc	test	40.0	46.4	50.0	53.4	36.8	46.7	49.8	52.0	**57.6**
BoolQ	acc	dev	56.6	62.7	72.2	**83.1**	56.6	56.1	73.6	78	60.5
Copa	acc	dev	73	85	86	90	67	80	86	90	**91**
RTE	acc	dev	45.8	58.8	60.3	**67.9**	51.3	49.1	63.8	50.5	63.5
WiC	acc	dev	50.0	49.8	49.5	50.3	**50.8**	50.3	44	50.6	0.0
Multirc	f1a	dev	57.7	58.0	52.4	**73.7**	58.6	53.0	39.0	54.8	72.9
WSC	acc	dev	65.6	79.3	81.8	**85.3**	66.3	77.2	80.7	82.8	65.4
ReCoRD	acc	dev	77.5	87.1	88.9	**90.3**	71.6	86.7	89.2	**90.3**	90.2
CB	acc	dev	**66.1**	33.9	40.7	48.2	42.9	37.5	33.9	42.9	46.4
ANLI R1	acc	dev	34.1	33.9	33.4	39.2	36.1	33.2	34.7	**39.4**	34.6
ANLI R2	acc	dev	33.8	32.4	34.9	**37.3**	36.7	33.6	34.8	35.7	35.4
ANLI R3	acc	dev	32.8	34.0	34.6	**41.3**	34.8	34.1	34.9	34.6	34.5
Avg NLG	-	-	18.6	35.1	39.6	**54.6**	14.9	31.3	38.0	45.8	47.6
Avg NLU	-	-	51.5	58.3	61.1	**66.2**	48.9	56.1	60.2	63.2	60.8

表13。在所有29个基准测试的GPT3和不同的Glam Moe和密集模型上的一声分数。

			GLaM (MoE)				GLaM (Dense)				GPT3
Name	Metric	Split	0.1B/64E	1.7B/64E	8B/64E	64B/64E	0.1B	1.7B	8B	137B	GPT-3 (175B)
TriviaQA	acc (em)	dev	15.2	54.1	65.9	**75.8**	8.3	36.3	56.4	70.0	68.0
NQs	acc (em)	test	2.5	10.7	16.0	**26.3**	1.19	6.5	10.7	19.1	23.0
WebQS	acc (em)	test	5.9	13.9	17.0	24.4	3.44	9.3	11.6	18.8	**25.3**
Lambada	acc (em)	test	36.9	57.4	64.1	**80.9**	21.8	52.3	64.7	68.5	72.5
HellaSwag	acc	dev	43.5	66.4	74.0	76.8	34.7	60.5	72.6	76.8	**78.1**
StoryCloze	acc	test	67.0	77.9	80.0	84.0	63.7	76.4	82.1	82.6	**84.7**
Winograd	acc	test	69.2	80.2	85.3	83.9	65.6	80.2	84	85.3	**89.7**
WinoGrande	acc	dev	51.7	63.5	68.7	73.0	49.8	62.8	70.0	73.1	**73.2**
DROP	f1	dev	16.3	24.8	28.4	**57.8**	19.3	24.9	41.2	49.4	34.3
CoQA	f1	dev	48.3	72.8	76	79.6	33.3	72.7	74.4	78.8	**84.0**
QuAC	f1	dev	28.7	35.2	43.1	42.7	23.7	35.7	35.1	**44.6**	43.4
SQuADv2	f1	dev	35.5	69.5	76.3	**71.8**	34.2	67.1	69.2	70.0	65.4
SQuADv2	acc (em)	dev	21.8	53.6	60.9	**66.5**	29.0	50.8	64.2	63.7	60.1
RACE-m	acc	test	42.7	60.9	60.6	65.5	43.1	56.4	63.1	**69.0**	57.4
RACE-h	acc	test	29.1	41.9	44.6	**48.7**	29.4	40.8	45.3	47.7	45.9
PIQA	acc	dev	69.0	76.0	78.1	**81.4**	63.7	73.1	76.3	79.5	80.5
ARC-e	acc	test	53.5	68.1	73.4	76.6	45.9	63.8	62.6	**77.2**	71.2
ARC-c	acc	test	27.0	39.3	44.8	50.3	24.5	35.2	41.5	50.7	**53.2**
Openbookqa	acc	test	39.6	47.6	50.6	55.2	37.8	47.2	53.0	55.4	**58.8**
BoolQ	acc	dev	53.6	62.0	70.8	**82.8**	55.7	58.1	76.4	77.5	76.7
Copa	acc	dev	75	81	86	**92**	71	81	86	91	87
RTE	acc	dev	53.1	54.5	57.0	**71.5**	53.4	55.2	62.0	58.4	70.4
WiC	acc	dev	47.3	47.0	48.0	**52.7**	47.3	46.8	48.0	48.7	48.6
Multirc	f1a	dev	58.5	59.6	62.0	**74.7**	56.3	59.4	61.9	64.2	72.9
WSC	acc	dev	67.7	77.5	83.8	83.9	63.8	78.5	83.0	**86.3**	69.2
ReCoRD	acc	dev	77.5	87.3	89.0	**90.3**	71.6	86.2	89.2	90.2	90.1
CB	acc	dev	41.1	35.7	44.6	**73.2**	42.9	41.1	30.4	48.2	64.3
ANLI R1	acc	dev	32.1	31.1	32.3	**42.4**	32.5	31.4	31.9	34.8	32.0
ANLI R2	acc	dev	31.1	30.7	32.5	**40.0**	30.7	31.2	30.7	32.6	33.9
ANLI R3	acc	dev	30.5	31.6	34.8	**40.8**	30.9	30.3	32.4	35.0	35.1
Avg NLG	-	-	23.5	43.6	49.7	**58.4**	19.4	39.5	47.5	52.8	52.7
Avg NLU	-	-	50.4	58.1	61.9	**68.6**	48.3	56.9	61.7	65.0	65.4

表14。在所有29个基准测试的GPT3和不同的Glam Moe和密集模型上的分数很少。我们调整了GPT3使用的每个任务中相应值的镜头数。

			GLaM (MoE)				**GLaM (Dense)**				**GPT3**
Name	Metric	Split	0.1B/64E	1.7B/64E	8B/64E	64B/64E	0.1B	1.7B	8B	137B	GPT-3 (175B)
TriviaQA	acc (em)	dev	21.7	60.1	67.7	**75.8**	8.3	38.8	56.4	70.0	71.2
NQs	acc (em)	test	5.3	17.7	24.4	**32.5**	1.50	9.0	20.1	27.9	29.9
WebQS	acc (em)	test	12.1	24.4	29.6	41.1	6.90	9.3	25.5	32.9	**41.5**
Lambada	acc (em)	test	36.9	64.3	79.0	**86.6**	21.8	63.0	77.1	84.2	86.4
HellaSwag	acc	dev	45.6	66.2	74.0	77.2	34.7	60.7	72.6	76.8	**79.3**
StoryCloze	acc	test	69.4	80.0	82.8	86.7	63.7	78.7	83.7	85.7	**87.7**
Winograd	acc	test	69.2	82.8	85.3	**88.6**	65.6	80.5	85.4	85.3	**88.6**
WinoGrande	acc	dev	52.6	66.2	71.4	**79.2**	49.8	64.2	72.3	76.6	77.7
DROP	f1	dev	23.5	37.0	40.0	**58.6**	19.3	41.4	49.4	49.4	36.5
CoQA	f1	dev	48.3	66.0	72	79.6	33.3	66.0	74.4	78.8	**85.0**
QuAC	f1	dev	26.0	34.2	43.1	42.8	23.7	34.3	35.1	37.2	**44.3**
SQuADv2	f1	dev	38.7	61.8	67.1	**71.8**	34.2	60.0	69.6	70.0	69.8
SQuADv2	acc (em)	dev	32.7	55.5	60.9	**67.0**	29.0	53.9	64.2	63.7	64.9
RACE-m	acc	test	41.8	53.6	60.6	**66.9**	43.1	56.5	56	65.1	58.1
RACE-h	acc	test	31.5	40.2	44.6	**49.3**	29.5	40.8	43	48.1	46.8
PIQA	acc	dev	69.0	76.1	78.1	81.8	64.2	73.1	77	80.8	**82.3**
ARC-e	acc	test	57.8	70.1	75.3	**78.9**	48.9	66.0	74	79.0	70.1
ARC-c	acc	test	29.7	38.3	45.5	52.0	24.8	35.2	41.5	45.7	51.5
Openbookqa	acc	test	41.6	49.6	53.0	63.0	37.8	54	54.0	58.8	**65.4**
BoolQ	acc	dev	53.6	62.0	70.5	**83.1**	59.9	63.1	76.4	80.5	77.5
Copa	acc	dev	75	82	88	**93.0**	71	83	92.0	91.0	92.0
RTE	acc	dev	53.1	54.5	60.0	**76.2**	54.9	55.2	64.0	63.9	72.9
WiC	acc	dev	49.4	51.3	53.3	**56.3**	51.9	50.9	50.0	53.6	55.3
Multirc	f1a	dev	58.5	59.7	62.0	**77.5**	56.3	59.4	61.5	68.1	74.8
WSC	acc	dev	67.7	80.4	83.8	85.6	65.6	80.0	82.0	**87.4**	75.0
ReCoRD	acc	dev	77.5	87.3	89.0	**90.6**	71.8	86.2	89.0	90.5	89.0
CB	acc	dev	43.0	53.6	60.7	**84.0**	42.9	55.4	58	53.6	82.1
ANLI R1	acc	dev	34.3	31.4	34.0	**44.3**	33.5	33.1	33.2	35.8	36.8
ANLI R2	acc	dev	32.3	33.0	32.0	**41.2**	34.4	33.7	33.9	35.6	34.0
ANLI R3	acc	dev	33.9	35.8	33.0	**44.7**	32.9	33.3	35.0	34.7	40.2
Avg NLG	-	-	27.2	46.8	53.0	**61.6**	19.8	42.7	52.4	57.1	58.8
Avg NLU	-	-	51.7	59.7	63.6	**71.4**	49.2	59.2	63.7	66.8	68.4