FINETUNED LANGUAGE MODELS ARE ZERO-SHOT LEARNERS

微调语言模型是零镜头学习者

# abstract

本文探索了一种提高语言模型零样本学习能力的简单方法。 我们展示了指令调优——在通过指令描述的一系列任务上微调语言模型——大大提高了在看不见的任务上的零样本性能。

我们采用一个 137B 参数的预训练语言模型，并在通过自然语言指令模板表达的 60 多个 NLP 任务上对其进行指令调整。 我们在看不见的任务类型上评估这个我们称之为 FLAN 的指令调整模型。 FLAN 大大提高了其未修改对应物的性能，并在我们评估的 25 项任务中的 19 项上超过了零样本 175B GPT-3。 FLAN 甚至在 ANLI、RTE、BoolQ、AI2-ARC、OpenbookQA 和 StoryCloze 上的性能都大大超过了少拍 GPT-3。 消融研究表明，任务数量和模型规模是指令调整成功的关键因素。

# 1 INTRODUCTION

大规模的语言模型 (LM)，例如 GPT-3 (Brown et al., 2020)，已被证明可以非常好地执行小样本学习。然而，他们在零样本学习方面不太成功。例如，在阅读理解、问答和自然语言推理等任务上，GPT-3 的零样本性能比少样本性能差很多。一个潜在的原因是，如果没有少样本样本，模型很难在与预训练数据格式不相似的提示上表现良好。

在本文中，我们探索了一种简单的方法来提高大型语言模型的零样本性能，从而将其影响范围扩大到更广泛的受众。我们利用 NLP 任务可以通过自然语言指令来描述的直觉，例如“这个电影评论的情绪是正面的还是负面的？”或“将‘你好吗’翻译成中文。”我们采用 137B 参数的预训练语言模型并执行指令调优——在通过自然语言指令表达的 60 多个 NLP 任务的混合上对模型进行微调。我们将这个结果模型称为 Finetuned LANguage Net，或 FLAN。

为了评估 FLAN 在未知任务上的零样本性能，我们根据 NLP 任务的任务类型将它们分组到集群中，并在对所有其他集群进行指令调整 FLAN 的同时，将每个集群用于评估。例如，如图 1 所示，为了评估 FLAN 执行自然语言推理的能力，我们在一系列其他 NLP 任务（如常识推理、翻译和情感分析）上对模型进行指令调整。由于此设置确保 FLAN 在指令调整中没有看到任何自然语言推理任务，因此我们评估了其执行零样本自然语言推理的能力。

我们的评估表明，FLAN 显着提高了基本 137B 参数模型的零样本性能。在我们评估的 25 个任务中的 19 个任务中，FLAN 的零样本也优于 175B 参数 GPT-3 的零样本，甚至在许多任务（如 ANLI、RTE、BoolQ、 AI2-ARC、OpenbookQA 和 StoryCloze。在消融研究中，我们发现在指令调整中增加任务集群的数量可以提高看不见的任务的性能，并且指令调整的好处只有在模型规模足够大的情况下才会出现。

我们的实证结果强调了语言模型执行使用自然语言指令描述的任务的能力。更广泛地说，如图 2 所示，指令调优结合了预训练-微调和提示范式的吸引人的特征，通过微调使用监督来提高语言模型响应推理时间文本交互的能力。用于加载用于 FLAN 的指令调整数据集的源代码在 https://github.com/google-research/flan 上公开提供。

# 2 FLAN: INSTRUCTION TUNING IMPROVES ZERO-SHOT LEARNING

指令调优的动机是提高语言模型响应 NLP 指令的能力。 这个想法是，通过使用监督来教 LM 执行通过指令描述的任务，它将学会遵循指令，甚至对看不见的任务也这样做。 为了评估模型在未知任务上的性能，我们按任务类型将任务分组到集群中，并保留每个任务集群进行评估，同时对所有剩余集群进行指令调整。

## 2.1 TASKS & TEMPLATES

从头开始创建具有大量任务的可行指令调整数据集将是资源密集型的。因此，我们选择将研究社区创建的现有数据集转换为教学格式。我们将 62 个在 Tensorflow Datasets 上公开可用的文本数据集（包括语言理解和语言生成任务）聚合到一个单一的混合物中。图 3 显示了我们使用的所有数据集；每个数据集被归类为十二个任务集群之一，给定集群中的数据集属于相同的任务类型。我们请相关读者参阅附录 B，了解单个数据集的描述、大小和示例。

我们将任务定义为由数据集给出的一组特定的输入-输出对（例如，我们认为 RTE 和 ANLI 是独立的任务，即使它们都是蕴涵）。对于每个任务，我们手动编写十个独特的模板，使用自然语言指令描述任务。十个模板中的大多数描述了原始任务，但为了增加多样性，对于每个任务，我们最多包含三个模板来“扭转任务”（例如，对于情感分类，我们包含要求生成负面电影评论的模板）。然后我们在所有任务的混合上对预训练的语言模型进行指令调整，每个任务中的示例通过为该任务随机选择的指令模板格式化。图 4 显示了用于自然语言推理任务的多个指令模板。

## 2.2 EVALUATION SPLITS

我们感兴趣的是 FLAN 如何在指令调整中未见过的任务上执行，因此定义什么算作看不见的任务至关重要。 虽然一些先前的工作通过禁止在训练中出现相同的数据集来对未见过的任务进行分类，但我们将使用更保守的定义，利用图 3 中的任务集群。在这项工作中，如果没有任务，我们只考虑在评估时看不见的任务 T 在指令调整期间可以看到 T 所属的任何集群。 例如，如果 T 是一个蕴涵任务，那么指令调整数据集中没有出现任何蕴涵任务，我们对来自所有其他集群的任务进行指令调整。 2 因此，使用这个定义来评估 FLAN 在跨越 c 的任务上的性能 集群，我们对 c 个集群间分裂的模型进行指令调整，其中每个分裂在指令调整期间提供一个不同的集群。

## 2.3 CLASSIFICATION WITH OPTIONS

给定任务所需的输出空间是几个给定类别之一（例如分类）或自由文本（例如生成）。 由于 FLAN 是仅解码器语言模型的指令调整版本，它自然会以自由文本形式响应，因此对于所需输出是自由文本的任务，无需进一步修改。

对于分类任务，先前的工作 (Brown et al., 2020) 使用了等级分类方法，例如，只考虑两个输出（“是”和“否”），并将较高概率的一个作为模型的预测 . 虽然这个程序在逻辑上是合理的，但它是不完美的，因为答案的概率质量可能在每个答案的表达方式之间具有不希望的分布（例如，说“是”的大量替代方式可能会降低分配给“的概率质量”） 是的”）。 因此，我们包含一个选项后缀，其中我们将标记 OPTIONS 与该任务的输出类列表一起附加到分类任务的末尾。 这使模型知道在响应分类任务时需要哪些选择。 选项的使用示例显示在 NLI 和图 1 中的常识示例中。

## 2.4 TRAINING DETAILS

模型架构和预训练。在我们的实验中，我们使用密集的从左到右、仅解码器的 137B 参数转换器语言模型。该模型使用 SentencePiece 库 (Kudo & Richardson, 2018) 在一组 Web 文档（包括那些带有计算机代码的文档）、对话数据和维基百科上进行预训练，这些文档被标记为 2.81T BPE 标记和 32K 标记的词汇表。大约 10% 的预训练数据是非英语的。这个数据集不像 GPT-3 训练集那么干净，而且还混合了对话和代码，因此我们预计这个预训练的 LM 在 NLP 任务上的零和少拍性能会略低。此后，我们将此预训练模型称为 Base LM。这个相同的模型以前也用于程序合成（Austin 等人，2021）。

指令调整程序。 FLAN 是 Base LM 的指令调整版本。我们的指令调优管道混合了所有数据集并从每个数据集中随机采样示例。一些数据集有超过一千万个训练示例（例如翻译），因此我们将每个数据集的训练示例数量限制为 30,000。其他数据集的训练示例很少（例如，CommitmentBank 只有 250 个），因此为了防止这些数据集被边缘化，我们遵循示例比例混合方案（Raffel 等人，2020），最大混合率为 3,000.3 我们微调使用 Adafactor Optimizer (Shazeer & Stern, 2018) 以 3e-5 的学习率以 8,192 的批量大小进行 30,000 次梯度更新的所有模型。我们的微调过程中使用的输入和目标序列长度分别为 1024 和 256。我们使用打包 (Raffel et al., 2020) 将多个训练示例组合成一个序列，使用特殊的序列结束标记将输入与目标分开。

# 3 RESULTS

在本节中，我们在自然语言推理、阅读理解、开放域 QA、常识推理、共指解析和翻译等任务上评估 FLAN。如第 2.2 节所述，我们通过将任务分组到集群中并保留每个集群进行评估，同时对所有剩余集群进行指令调整来评估未见过的任务。对于每项任务，我们报告所有模板的性能均值和标准差，这代表了给定典型自然语言指令的 FLAN 的预期性能。由于开发集有时可用于手动提示工程（Brown 等人，2020 年），对于每项任务，我们还使用具有最佳开发集性能的模板报告测试集性能。

为了比较，我们包括了 175B GPT-3 的零次/少次性能。使用最好的开发模板，零样本 FLAN 在 25 项任务中的 19 项上优于零样本 GPT-3，甚至在 10 项任务上超过 GPT-3 的少样本性能。我们还使用与 GPT-3 相同的提示从 Base LM 报告零和少拍4结果，因为 Base LM 不适合没有指令调整的自然指令。该基线提供了指令调整有多大帮助的最直接消融。指令调优显着改善了大多数任务的 Base LM。在监督模型方面，我们在可用时报告了 T5-11B (Raffel et al., 2020) 的性能；否则，我们将展示 BERT-large 的性能（Devlin 等人，2019 年）。

总的来说，我们观察到指令调优对于可以自然表达为指令的任务（例如，自然语言推理和问答）非常有效，而对于直接表述为语言建模的任务（例如，常识推理和共指解析）则不太有效）。

## 3.1 NATURAL LANGUAGE INFERENCE

表 1 显示了自然语言推理 (NLI) 的结果，其中给定前提和假设，模型必须在前提为真的情况下确定假设是否为真。 FLAN 在所有情况下都具有强大的性能。 尽管 CB (de Marneffe et al., 2019) 和 RTE (Wang et al., 2018) 的不同模板的结果之间存在很大差异，但 FLAN 在没有任何提示的情况下，在四个数据集上大大优于零样本 GPT-3 工程。 使用最好的开发模板，FLAN 在所有五个数据集上的表现都优于小样本 GPT-3。 FLAN 甚至超过了 ANLI-R3 上的监督 BERT（Nie 等人，2020）。

正如布朗等人所指出的。 （2020），GPT-3 与 NLI 斗争的一个原因可能是 NLI 示例不太可能自然地出现在无监督的训练集中，因此被笨拙地措辞为句子的延续。 另一方面，对于 FLAN，我们将 NLI 表述为更自然的问题“是否意味着那个？” 并获得更高的性能。

## 3.2 READING COMPREHENSION & OPEN-DOMAIN QA

阅读理解的结果显示在表 2 中，其中要求模型回答有关所提供文章的问题。 FLAN 在 BoolQ（Clark 等人，2019a）和 OBQA（Mihaylov 等人，2019 年）上大大优于 GPT-3。 , 2018)，尽管 Base LM 已经在 BoolQ 上实现了高性能。在 MultiRC (Khashabi et al., 2018) 上，当使用最好的开发模板时，FLAN 的性能略高于少拍 GPT-3。

表 2 中还显示了开放域问答的结果，它要求模型回答有关世界的问题，而无需访问包含答案的特定信息。在 ARC-easy 和 ARC-challenge (Clark et al., 2018) 上，FLAN 的性能大大高于零样本和少样本 GPT-3。对于自然问题（NQ；Lee 等人，2019 年；Kwiatkowski 等人，2019 年）FLAN 优于零样本 GPT-3，但不优于少样本 GPT-3。然而，对于 TriviaQA (Joshi et al., 2017)，FLAN 的性能低于零样本 GPT-3，后者实现了卓越的性能并击败了受监督的 T5-11B。

## 3.3 COMMONSENSE REASONING & COREFERENCE RESOLUTION

五个常识推理数据集的结果如表 3 所示。 FLAN 在 StoryCloze（Mostafazadeh 等人，2016 年）上的性能优于 GPT-3，并且在 CoPA（Roemmele 等人，2011 年）和 PiQA（Bisk 等人，2020 年）上的性能相当）。然而，对于 HellaSwag（Zellers 等人，2019）和 ReCoRD（Zhang 等人，2018 年），Base LM 和 FLAN 的表现都远低于 GPT-3。

表 3 还显示了两个共指分辨率数据集的结果。具有最佳开发模板的 FLAN 在 Winogrande 上的性能优于零样本 GPT-3（Sakaguchi 等人，2020 年）。对于 WSC273（Levesque 等人，2012），Base LM 和 FLAN 都远低于 GPT-3（我们无法计算 FLAN 的最佳开发模板结果，因为 WSC273 没有验证集）。

一个观察结果是，HellaSwag、PiQA、ReCoRD、WSC273 和 Winogrande 是直接在语言建模的原始预训练目标中制定的。因此，GPT-3 大大优于 FLAN 的结果与直觉一致，即当指令对于描述给定任务并不重要时，从 FLAN 获得的收益将是微不足道的。此外，我们注意到 FLAN 对这五种语言建模任务的进一步限制，包括选项实际上会损害性能，因此报告的结果是针对没有选项的等级分类。

## 3.4 TRANSLATION

与 GPT-3 类似，Base LM 的训练数据大约 90% 是英语，并包括一些其他语言的文本，这些文本并未专门用于训练模型执行机器翻译。我们还评估了 FLAN 在 GPT-3 论文中评估的三个数据集的机器翻译性能：来自 WMT'14 的法语-英语（Bojar 等人，2014）和来自 WMT'16 的德语-英语和罗马尼亚语-英语（Bojar等，2016）。

这些结果如表 4 所示。对于 Base LM，零镜头翻译性能较低，但少镜头翻译结果与 GPT-3 相当；在六项评估中的五项中，FLAN 的表现优于小样本 Base LM。与 GPT-3 相比，FLAN 在所有六项评估中都优于零样本 GPT-3，但在大多数情况下它的表现不如少样本 GPT-3。与 GPT-3 类似，FLAN 在翻译成英语方面表现出很强的结果，并且与有监督的翻译基线相比具有优势。然而，从英语翻译成其他语言的能力相对较弱，考虑到 FLAN 使用英语句子分词器并且大多数预训练数据是英语，这可能是预料之中的。

# 4 ABLATION STUDIES & FURTHER ANALYSIS

## 4.1 NUMBER OF INSTRUCTION TUNING CLUSTERS

由于我们论文的核心问题是指令调优如何提高模型在未知任务上的零样本性能，因此在第一次消融中，我们研究了指令调优中使用的集群和任务数量对性能的影响。对于此设置，我们将 NLI、开放域 QA 和常识推理作为评估集群，并使用剩余的 7 个集群进行指令调整。5 我们展示了 1-7 个指令调整集群的结果，其中集群以降序添加每个集群的任务数。

图 5 显示了这些结果。正如预期的那样，我们观察到三个保留集群的平均性能随着我们向指令调整（情感分析集群除外）添加额外的集群和任务而提高，证实了我们提出的指令调整方法对零样本的好处新任务的表现。更有趣的是，对于我们测试的七个集群，性能似乎没有饱和，这意味着性能可能会随着指令调整中添加更多集群而进一步提高。值得注意的是，尽管我们从情感分析集群中看到了最小的附加值，但这种消融不允许我们得出关于哪个指令调整集群对每个评估集群贡献最大的结论。

## 4.2 SCALING LAWS

正如布朗等人。 (2020) 表明语言模型的零样本和少样本能力对于较大的模型有显着提高，我们接下来探讨指令调整的好处如何受模型规模的影响。使用与之前的消融研究相同的集群拆分，我们评估了指令调整对大小为 422M、2B、8B、68B 和 137B 参数的模型的影响。

图 6 显示了这些结果。首先，作为对正确性的检查，我们在图 6A 中确认，对于所有模型大小，在指令调整期间看到的任务的性能比没有指令调整的未调整模型有很大的提高，正如人们所期望的那样。至于保留任务，在图 6B 中，我们看到对于 100B 参数数量级的两个模型，指令调整显着提高了性能，正如我们论文中的先前结果所预期的那样。然而，对于 8B 和更小的模型，在保持任务上的行为是发人深省的——指令调整实际上会损害保持任务的性能。对这一结果的一个潜在解释可能是，对于小规模模型，学习指令调整期间使用的 40 个任务会填满整个模型容量，导致这些模型在新任务上的表现更差。对于更大规模的模型，指令调整填充了一些模型容量，但也教会了这些模型遵循指令的能力，允许它们利用剩余容量泛化到新任务。

## 4.3 INSTRUCTION TUNING FACILITATES PROMPT TUNING

正如我们所看到的，指令调整提高了模型响应指令的能力，因此，如果 FLAN 确实更适合执行 NLP 任务，那么它也应该在使用通过提示优化的连续提示执行推理时获得更好的性能调音（Li & Liang，2021 年；Lester 等人，2021 年）。作为进一步分析，我们根据第 2.2 节中的集群拆分为每个 SuperGLUE (Wang et al., 2019a) 任务训练连续提示，这样在对任务 T 进行提示调整时，与 T 相同集群中的任务没有在指令调整期间看到。我们的即时调整设置遵循 Lester 等人的程序。 (2021) 除了我们使用 10 的提示长度，1e-4 的权重衰减，并且没有在注意力分数上使用 dropout；我们在初步实验中发现，这些变化提高了 Base LM 的性能。

表 5 显示了使用全监督训练集和仅具有 32 个训练示例的低资源设置的这些快速调整实验的结果。我们看到，在所有场景中，FLAN 的快速调优效果比 Base LM 更好。在很多情况下，特别是对于低资源设置，FLAN 上的即时调优甚至比 Base LM 上的即时调优提高了 10% 以上。这个结果以另一种方式举例说明了指令调整如何产生一个更适合执行 NLP 任务的检查点。

# 5 RELATED WORK

语言模型和多任务学习。我们的工作很大程度上受到了 NLP 应用程序语言模型的一系列先前工作的启发（Dai & Le，2015；Peters 等人，2018 年；Howard 和 Ruder，2018 年；Radford 等人，2018 年；2019 年等） .指令调整可以看作是多任务学习 (MTL) 的一种表述，它是深度学习中的一个既定领域（Collobert 等人，2011 年；Luong 等人，2016 年；Ruder，2017 年；Velay & Daniel，2018 年；Clark 等人al., 2019b; Liu et al., 2019b，除其他外）—参见 Worsham & Kalita (2020)，了解最近对 NLP 的 MTL 的调查。与之前的 MTL 工作侧重于跨训练任务的性能改进（Raffel 等人，2020 年；Aghajanyan 等人，2021 年）或新领域（Axelrod 等人，2011 年）不同，我们的工作侧重于零样本泛化到训练中没有看到的任务。

零镜头学习。我们的工作也属于完善的零样本学习类别，它历来被用来指在一组看不见的类别中对实例进行分类（Lampert 等人，2009 年；Romera-Paredes 和 Torr，2015 年；Srivastava 等人al.，2018 年；Yin 等人，2019 年等）。在 NLP 中，零样本学习工作还包括在看不见的语言对之间进行翻译（Johnson 等人，2017 年；Pham 等人，2019 年）、对看不见的语言进行语言建模（Lauscher 等人，2020 年）以及各种 NLP应用（Liu 等，2019a；Corazza 等，2020；Wang 等，2021；Reif 等，2021）。最近，语言模型的新兴能力（Brown 等人，2020 年）引起了人们对模型如何泛化到看不见的任务的兴趣，即我们论文中使用的零样本学习的定义。此外，元学习（Finn 等人，2017 年；Vanschoren，2018 年等）也广泛地尝试训练快速适应未知任务的模型。

提示。指令调整利用了大规模语言模型包含大量世界知识并可以执行一系列 NLP 任务的直觉（Brown 等人，2020，另见 Bommasani 等人（2021））。共享此目标的另一行工作提示具有通过反向传播优化的连续输入的模型，以显着提高性能（Li 和 Liang，2021 年；Lester 等人，2021 年；Qin 和 Eisner，2021 年）。尽管这些方法的成功在很大程度上取决于模型规模（Lester 等人，2021），大型模型的服务成本可能很高，但单个大型模型执行多项任务的能力稍微减轻了这种负担。

多任务问答。我们用于指令调优的指令类似于 McCann 等人的多任务问答工作。 (2018)，它将 10 个 NLP 任务作为上下文的问答，并在用自然语言提示制定的一系列任务上训练模型。他们报告了微调任务的迁移学习收益以及 SNLI（Bowman 等人，2015 年）和 Amazon/Yelp 评论（Kotzias 等人，2015 年）上的零样本域适应结果。虽然麦肯等人。 (2018) 没有利用无监督的预训练，只报告零样本转移到看不见的领域，我们的工作使用预训练的 LM 并专注于看不见的任务集群的零样本性能。专注于二进制文本分类，Zhong 等人。 (2021) 在 43 个任务上微调 T5-770M，这些任务被表述为是/否问题，并研究了在看不见的任务上的零样本性能。相比之下，我们的论文范围要大得多，用更大的模型在广泛的任务上凭经验证明了这个想法。

基于指令的 NLP。最近语言模型能力的改进导致人们对基于指令的 NLP 新兴领域的兴趣增加（Goldwasser & Roth，2014 年，参见 McCarthy（1960））。 Schick & Schütze (2021) 使用完形填空式短语中的任务描述来帮助语言模型为半监督学习分配软标签。 Efrat & Levy (2020) 在从检索句子的第 n 个单词到为 SQuAD 生成示例等简单任务上评估了 GPT-2（Radford 等人，2019），得出的结论是 GPT-2 在所有任务中表现不佳。米什拉等人。 (2021) 使用诸如问答、文本分类和文本修改等任务的指令微调 BART（Lewis 等人，2020）。除了模型规模和评估任务数量不同之外，我们的工作与他们的工作有很大不同，因为我们研究了对新任务的泛化。而 Mishra 等人。 (2021) 对相似的任务分布进行微调和评估，我们在本文中使用的集群间分割评估了模型在微调期间没有看到类似任务时执行新任务的能力。叶等人。 (2021) 介绍了跨任务小样本学习的设置，发现多任务学习提高了 BART 在未知下游任务上的小样本能力。最后，OpenAI 有一个名为 Instruct 系列的测试版产品。虽然他们目前还没有发布关于它的研究论文，但根据他们网站上发布的例子6，他们的产品似乎与我们的工作目标相似。

# 6 DISCUSSION

我们的论文探讨了零样本提示中的一个简单问题：指令调整语言模型是否会提高其执行未知任务的能力？我们在 FLAN 上的实验表明，指令调优提高了针对未调优模型的性能，并在我们评估的大多数任务上超过了零样本 GPT-3。通过消融研究，我们了解到，随着指令调整中使用的任务集群数量的增加，看不见的任务的性能会提高，而且有趣的是，指令调整的好处只有在模型规模足够大的情况下才会出现。此外，我们的 FLAN 似乎比未修改的基本模型对即时调整的响应更好，证明了指令调整的额外好处。

我们研究中的一个限制是将任务分组到集群中存在一定程度的主观性（例如，情感分析可以被视为阅读理解的一个小子集），因为没有公认的方法来操作两个任务之间的相似性.因此，我们根据文献中公认的分类将任务分配给集群，当任务似乎属于多个集群时采取保守的方法（例如，在评估阅读理解和常识推理时，从指令调整中排除带有常识的阅读理解）。作为另一个限制，我们使用简短的指令（通常是一个句子）来描述众所周知的 NLP 任务。其他任务可能需要更长或更具体的说明来充分描述，以及涉及示例的解释；我们将这些场景留给未来的工作。

本文中显示的结果为未来的研究提出了几个有希望的方向。尽管 FLAN 对 60 多个数据集进行了指令调整，但这些数据集仅涵盖 10 个任务集群（加上一些杂项任务），考虑到此类模型可用于所有潜在任务的数量相对较少。例如，可以通过更多指令调整任务来进一步提高性能，这些任务可以以自我监督的方式生成。除了收集更多任务之外，探索多语言环境也很有价值，例如，人们可以问，在高资源语言中对监督数据进行指令调整是否会提高低资源语言中新任务的性能。最后，具有监督数据的指令调整模型也可能用于改善模型在偏见和公平方面的行为（Solaiman & Dennison，2021）。

# 7 CONCLUSIONS

本文探讨了指令调优。我们提出了 FLAN，这是一个 137B 参数语言模型，它执行使用指令描述的 NLP 任务。通过利用微调的监督来提高语言模型响应教学提示的能力，FLAN 结合了预训练-微调和提示范式的吸引人的方面。 FLAN 的性能优于零样本和少样本 GPT-3，表明大规模模型遵循指令的潜在能力。我们希望我们的论文能够推动对零样本学习和使用标记数据改进语言模型的进一步研究。

道德考虑

这项工作使用语言模型，Bender & Koller (2020)、Brown 等人讨论了该模型的风险和潜在危害。 (2020)，Bender 等人。 (2021)、Patterson 等人 (2021) 等。由于我们在本文中的贡献不是预训练的语言模型本身，而是关于指令调整如何影响语言模型在未知任务上的零样本性能的实证研究，因此我们还强调了几个相关的伦理考虑。首先，标记数据集（例如我们用于微调的数据集）可能包含不良偏差，并且这些偏差可以传播到模型在下游任务上的零样本应用程序中。其次，指令调优模型可能需要更少的数据和专业知识来使用；这种较低的准入门槛可能会增加此类模式的好处和相关风险。最后，我们要明确的是，对于大多数数据集，BERT 和 T5 等监督模型仍然优于本文中的仅零样本指令模型。

环境考虑

我们使用与 Austin 等人相同的预训练语言模型。 (2021)。这些预训练模型的能源成本和碳足迹分别为 451 MWh 和 26 tCO2e。用于微调 flan 的额外指令调整梯度步数不到预训练步数的 2%，因此估计的额外能量成本相对较小。