
语言模型是无监督的多任务学习者

亚历克·雷德福德 杰弗里·吴 雷文·查尔德 大卫·鲁安 达里奥·阿莫迪 伊利亚·苏茨克弗

抽象的

自然语言处理任务，例如问答、机器翻译、阅读理解和摘要，通常采用监督学习的方式，在特定任务的数据集上进行处理。我们证明，当语言模型在名为 WebText 的包含数百万个网页的新数据集上进行训练时，无需任何明确的监督即可开始学习这些任务。当以文档和问题为条件时，该语言模型生成的答案在 CoQA 数据集上的 F1 值达到 55，在不使用超过 127,000 个训练样本的情况下，达到或超过了 4 个基线系统中的 3 个的性能。语言模型的容量对于零样本任务迁移的成功至关重要，提高容量可以以对数线性的方式提高跨任务的性能。我们最大的模型 GPT-2 是一个拥有 1.5B 参数的 Transformer，它在 8 个测试语言建模数据集上的 7 个上取得了最佳结果，但在 WebText 上仍然欠拟合。来自该模型的样本反映了这些改进，并且包含连贯的文本段落。这些发现为构建能够从自然发生的演示中学习执行任务的语言处理系统提供了一条有希望的途径。

称职的通才。我们希望开发出能够执行多种任务的通用系统，最终无需为每个任务手动创建和标记训练数据集。

创建机器学习系统的主要方法是收集一个训练样本数据集，其中包含展示目标任务正确行为的样本，训练一个系统来模仿这些行为，然后在独立同分布 (IID) 的保留样本上测试其性能。这种方法在狭义专家模型上取得了良好的进展。但是，字幕模型 (Lake et al., 2017)、阅读理解系统 (Jia & Liang, 2017) 和图像分类器 (Alcorn et al., 2018) 在输入的多样性和多样性方面往往表现不稳定，凸显了这种方法的一些缺陷。

我们怀疑，在单一领域数据集上进行单任务训练的盛行是导致当前系统泛化能力不足的主要原因。要构建基于当前架构的鲁棒系统，可能需要在广泛的领域和任务上进行训练和性能测量。最近，一些基准测试（例如 GLUE (Wang et al., 2018) 和 decaNLP (McCann et al., 2018)) 已开始对此进行研究。

1. 简介

如今，机器学习系统通过使用海量数据集、高容量模型和监督学习 (Krizhevsky 等人, 2012) (Sutskever 等人, 2014) (Amodei 等人, 2016) 的组合，在它们所训练的任务上表现出色（这是预期）。然而，这些系统很脆弱，对数据分布 (Recht 等人, 2018) 和任务规范 (Kirkpatrick 等人, 2017) 的细微变化很敏感。当前的系统更适合被描述为“狭隘的专家”，而不是

多任务学习 (Caruana, 1997) 是一个有望提升通用性能的框架。然而，自然语言处理中的多任务训练仍处于起步阶段。最近的研究报告了适度的性能提升 (Yogatama 等人, 2019)，迄今为止最雄心勃勃的两项研究分别在总共 10 个和 17 个（数据集，目标）对上进行了训练 (McCann 等人, 2018) (Bowman 等人, 2018)。从元学习的角度来看，每个（数据集，目标）对都是从数据集和目标的分布中采样的单个训练示例。当前的机器学习系统需要数百到数千个示例才能导出具有良好泛化能力的函数。这表明，多任务训练可能需要同样多的有效训练对才能在当前方法下实现其潜力。继续扩展数据集的创建和目标的设计将非常困难，以至于我们可能需要使用现有技术强行实现这一目标。这促使人们探索执行多任务学习的额外设置。

同等贡献者：OpenAI，美国加利福尼亚州旧金山。联系人：Alec Radford。

目前在语言任务上表现最好的系统

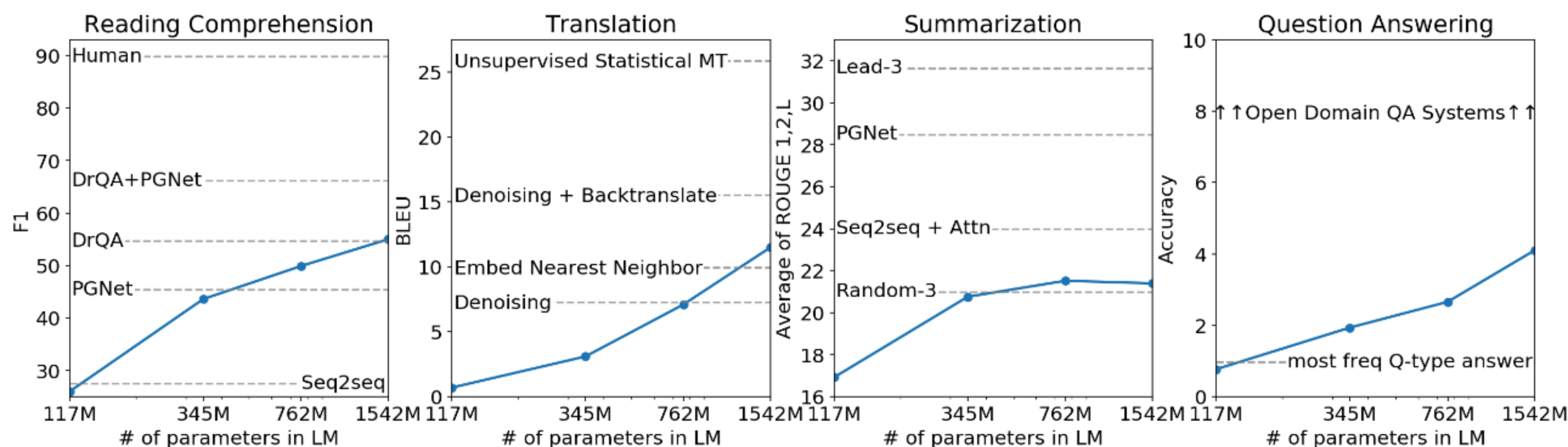


图 1. WebText 语言模型在多种自然语言处理任务中的零样本任务性能随模型大小的变化。阅读理解结果分别来自 CoQA (Reddy 等人, 2018)、WMT-14 Fr-En 翻译 (Artetxe 等人, 2017)、CNN 和《每日邮报》摘要 (See 等人, 2017) 以及自然问答 (Kwiatkowski 等人, 2019) 的测试。第 3 节包含每个结果的详细描述。

利用预训练和监督微调相结合的方法。这种方法历史悠久，且趋向于更灵活的迁移形式。首先，学习词向量并将其用作特定任务架构的输入 (Mikolov et al., 2013) (Collobert et al., 2011)，然后迁移循环网络的上下文表征 (Dai & Le, 2015) (Peters et al., 2018)。最近的研究表明，特定任务架构不再必要，迁移多个自注意力模块就足够了 (Radford et al., 2018) (Devlin et al., 2018)。

这些方法仍然需要监督训练才能执行任务。当只有极少或没有监督数据可用时，另一项研究已经证明了语言模型在执行特定任务方面的潜力，例如常识推理 (Schwartz 等人, 2017) 和情感分析 (Radford 等人, 2017)。

在本文中，我们将这两条工作线连接起来，并延续了迁移方法更通用化的趋势。我们证明了语言模型可以在零样本设置下执行下游任务，而无需任何参数或架构修改。我们通过强调语言模型在零样本设置下执行各种任务的能力，证明了这种方法的潜力。根据任务的不同，我们取得了有希望的、有竞争力的、甚至是最先进的结果。

符号作为条件概率的乘积 (Jelinek & Mercer, 1980) (Bengio et al., 2003):

$$p(x) = \prod_{i=1}^T p(s_i | s_{1:i-1}, s) \quad (1)$$

这种方法能够轻松地从 $p(x)$ 以及任何形式为 $p(s_{1:T} | s_{1:T-1}, s)$ 的条件句中进行采样和估计。近年来，能够计算这些条件概率的模型的表达能力得到了显著提升，例如 Transformer 等自注意力架构 (Vaswani 等人, 2017)。

在概率框架中，学习执行单个任务可以表示为估计条件分布 $p(\text{输出} | \text{输入})$ 。由于通用系统应该能够执行许多不同的任务，即使针对相同的输入，它也应该不仅以输入为条件，还要以要执行的任务为条件。也就是说，它应该对 $p(\text{输出} | \text{输入}, \text{任务})$ 进行建模。这在多任务和元学习环境中已被各种形式化。任务条件通常在架构级别实现，例如 (Kaiser et al., 2017) 中针对特定任务的编码器和解码器，或在算法级别实现，例如 MAML 的内外循环优化框架 (Finn et al., 2017)。但正如 McCann et al. (2018) 所举例说明的那样，语言提供了一种灵活的方式来将任务、输入和输出全部指定为符号序列。例如，翻译训练示例可以写成序列 (翻译成法语、英语文本、法语文本)。同样，阅读理解训练示例可以

2. 方法

我们方法的核心是语言建模。语言建模通常被定义为对一组示例 (x_1, x_2, \dots, x_n) 进行无监督分布估计，每个示例由可变长度的符号序列 (s_1, s_2, \dots, s_n) 组成。由于语言具有自然的顺序性，因此通常将联合概率分解为

写成 (回答问题，记录，问题，答案)。McCann 等人 (2018) 证明了可以训练一个单一模型，即 MQAN，

对具有这种格式的示例进行推断和执行许多不同的任务。

语言建模原则上也能够学习 McCann 等人 (2018) 的任务，而无需明确监督哪些符号是需要预测的输出。由于监督目标与无监督目标相同，但仅在序列的子集上进行评估，因此无监督目标的全局最小值也是监督目标的全局最小值。在这个略显玩具化的设置中，(Sutskever et al., 2015) 中讨论的将密度估计作为原则性训练目标的担忧被回避了。问题变成了我们是否能够在实践中优化无监督目标以使其收敛。初步实验证实，足够大的语言模型能够在这个类似玩具的设置中执行多任务学习，但学习速度比明确监督的方法要慢得多。虽然从上述设定设置到“自然语言”的混乱状态还有很长的路要走，但 Weston (2016) 在对话的背景下主张开发能够直接从自然语言中学习的系统，并展示了一个概念证明——通过对教师输出的前向预测，学习无需奖励信号的问答任务。虽然对话是一种有吸引力的方法，但我们担心它过于局限。互联网包含大量无需交互式交流即可被动获取的信息。我们推测，无论获取信息的方法如何，具有足够容量的语言模型将开始学习推断和执行自然语言序列中演示的任务，以便更好地预测它们。如果语言模型能够做到这一点，它实际上就是在进行无监督的多任务学习。我们通过分析语言模型在零样本设置下对各种任务的表现来测试情况是否如此。

2.1. 训练数据集

先前的研究大多基于单一文本领域训练语言模型，例如新闻文章 (Jozefowicz 等人, 2016)、维基百科 (Merity 等人, 2016) 或小说 (Kiros 等人, 2015)。我们的方法旨在构建尽可能庞大且多样化的数据集，以便收集尽可能多样化领域和情境中自然语言任务的演示。像 Common Crawl 这样的网络抓取工具是一个极具潜力的多样化且几乎无限的文本来源。虽然这些档案比当前的语言模型数据集大很多个数量级，但它们存在严重的数据质量问题。Trinh & Le (2018) 在其常识推理研究中使用了 Common Crawl，但指出大量文档“内容大多难以理解”。

我们在最初的实验中观察到了类似的数据问题，“我不是世界上最聪明的人，但就像法语里说的：Je ne suis pas un imbecile [我不是傻瓜]”。在 8 月 16 日发布的一篇现已删除的帖子中，若利埃特选区的保守党候选人索希尔·艾德 (Soheil Eid) 用法语写道：“Mentez mentez, il en restera toujours quelque chose”，翻译过来就是“撒谎，撒谎，总有一些东西会留下来。”

“我讨厌‘香水’这个词，”伯尔说，“法语里‘parfum’这个词会好一些。”

如果在 29:55 处仔细听，可以听到两个人用法语进行的对话：“-Comment on fait pour aller de l'autre coté? -Quel autre coté?”，意思是“- 你如何到达另一边? - 哪一边?”。

如果这听起来有点夸张，请考虑这个问题- 法语中的意思是：As-tu aller au cinéma?, 或者你去看电影了吗? 字面意思是你去看电影/剧院了吗?

“Brevet Sans Garantie Du Gouvernement”，翻译成英语：“已获得专利，无政府保证”。

表 1. 在整个 WebText 训练集中发现的英语到法语和法语到英语的翻译的自然发生的演示示例。

常见爬虫。Trinh & Le (2018) 的最佳成果是使用常见爬虫的一小分子样本取得的，该子样本仅包含与其目标数据集 Winograd Schema Challenge 最相似的文档。虽然这是一种提高特定任务性能的实用方法，但我们希望避免提前对要执行的任务做出假设。

相反，我们创建了一个新的网页抓取工具，重点关注文档质量。为此，我们只抓取经过人工筛选/整理的网页。手动筛选完整的网页抓取成本极高，因此，我们首先抓取了社交媒体平台 Reddit 上所有至少获得 3 个 Karma 的出站链接。这可以被视为一个启发式指标，用于判断其他用户认为该链接有趣、具有教育意义还是仅仅好笑。

最终的数据集 WebText 包含这 4500 万个链接的文本子集。为了从 HTML 响应中提取文本，我们结合使用了 Dragnet (Peters & Lecocq, 2013) 和 Newspapercontent 提取器。本文中呈现的所有结果均基于 WebText 的初步版本，该版本不包含 2017 年 12 月之后创建的链接，经过去重和一些基于启发式算法的清理后，包含略多于 800 万个文档，总计 40 GB 的文本。我们从 WebText 中删除了所有维基百科文档，因为它是其他数据集的通用数据源，并且由于数据量过大可能会使分析复杂化。

¹<https://github.com/codelucas/newspaper>

将训练数据与测试评估任务相结合。

2.2. 输入表示

通用语言模型 (LM) 应该能够计算 (并生成) 任何字符串的概率。当前的大规模语言模型包含预处理步骤, 例如小写化、分词和词汇外标记, 这些步骤限制了可建模字符串的空间。虽然将 Unicode 字符串处理为 UTF-8 字节序列可以优雅地满足这一要求, 正如 Gillick 等人 (2015) 的论文中所举例说明的那样, 但当前的字节级语言模型在大规模数据集 (例如十亿词基准测试 (AI-Rfou et al., 2018)) 上无法与词级语言模型竞争。我们在 WebText 上训练标准字节级语言模型的尝试中也观察到了类似的性能差距。

字节对编码 (BPE) (Sennrich 等人, 2015) 是一种介于字符级和词级语言建模之间的实用中间方法, 它有效地在频繁符号序列的词级输入和不频繁符号序列的字符级输入之间进行插值。尽管名称如此, 但参考的 BPE 实现通常操作的是 Unicode 码位而非字节序列。这些实现需要包含所有 Unicode 符号才能对所有 Unicode 字符串进行建模。这将导致在添加任何多符号标记之前, 基础词汇量超过 130,000 个。与 BPE 常用的 32,000 到 64,000 个标记词汇量相比, 这个规模过大。相比之下, 字节级版本的 BPE 只需要 256 大小的基础词汇量。然而, 由于 BPE 使用基于贪婪频率的启发式算法来构建标记词汇量, 因此将 BPE 直接应用于字节序列会导致合并效果不佳。我们观察到, BPE 会包含许多常见词汇 (例如 dog) 的变体, 因为它们有多种变体, 例如 dog. dog! dog?。这导致有限的词汇槽和模型容量分配不理想。为了避免这种情况, 我们阻止 BPE 合并任何字节序列的跨字符类别。我们添加了空格例外, 这显著提高了压缩效率, 同时在多个词汇标记之间仅增加了极小的单词碎片。这种输入表示使我们能够将词级语言模型的经验优势与字节级方法的通用性结合起来。由于我们的方法可以为任何 Unicode 字符串分配概率, 因此我们能够在任何数据集上评估我们的语言模型, 而无需考虑预处理、分词或词汇量。

2.3. 模型

我们的语言模型 (LM) 采用基于 Transformer (Vaswani 等人, 2017) 的架构。该模型在很大程度上遵循了 OpenAI GPT 模型 (Radford 等人, 2018) 的细节, 并采用了

| 参数层 d |
|----------------|
| 1.17 亿 12768 |
| 345 亿 24 1024 |
| 762 米 36 1280 |
| 1542 米 48 1600 |

表 2. 4 种模型尺寸的架构超参数。

进行了一些修改。层归一化 (Ba et al., 2016) 被移至每个子块的输入, 类似于预激活残差网络 (He et al., 2016), 并在最后一个自注意力模块之后添加了一个额外的层归一化。使用了一种改进的初始化方法, 该初始化方法考虑了残差路径随模型深度的累积。我们将初始化时残差层的权重缩放为 $1/\sqrt{N}$

其中 N 是残差层的数量。词汇量扩展到 50,257。我们还将上下文大小从 512 个 token 增加到 1024 个 token, 并使用了更大的 512 的批处理大小。

3.实验

我们训练并测试了四个大小近似对数均匀分布的语言模型 (LM)。表 2 总结了这些架构。最小的模型相当于原始 GPT, 第二小的模型相当于 BERT (Devlin et al., 2018) 中最大的模型。我们最大的模型, 我们称之为 GPT-2, 其参数比 GPT 多一个数量级。每个模型的学习率都经过手动调整, 以便在 5% 的 WebText 保留样本上获得最佳困惑度。所有模型仍然对 WebText 欠拟合, 并且随着训练时间的增加, 保留样本困惑度有所改善。

3.1. 语言建模

作为迈向零样本任务迁移的第一步, 我们感兴趣的是了解 WebText 语言模型 (LM) 在零样本领域迁移中的表现, 它主要针对语言建模进行训练。由于我们的模型在字节级别运行, 不需要有损预处理或分词, 因此我们可以在任何语言模型基准上对其进行评估。语言建模数据集的结果通常以每个典型预测单元 (通常是一个字符、一个字节或一个单词) 的平均负对数概率的缩放或指数形式来报告。我们通过根据 WebText 语言模型计算数据集的对数概率, 然后除以典型单元的数量来评估相同的数值。对于许多这样的数据集, WebText 语言模型的测试结果会明显偏离分布, 因为它们必须预测高度标准化的文本、分词伪像 (例如不连贯的标点符号和缩写)、打乱的句子, 甚至字符串

| | LAMBADA | LAMBADA | CBT-CN | CBT-NE | WikiText2 | PTB | enwik8 | text8 | WikiText103 | 1BW |
|---------|---------|---------|--------|--------|-----------|-------|--------|-------|-------------|--------|
| | (PPL) | (ACC) | (ACC) | (ACC) | (PPL) | (PPL) | (BPB) | (BPC) | (PPL) | (PPL) |
| 低于 | 99.8 | 59.23 | 85.7 | 82.3 | 39.14 | 46.54 | 0.99 | 1.08 | 18.3 | 21.8 |
| 117 百万 | 35.13 | 45.99 | 87.65 | 83.4 | 29.41 | 65.85 | 1.16 | 1.17 | 37.50 | 75.20 |
| 345 百万 | 15.60 | 55.48 | 92.35 | 87.1 | 22.76 | 47.33 | 1.01 | 1.06 | 26.37 | 55.72 |
| 762 百万 | 10.87 | 60.12 | 93.45 | 88.0 | 19.93 | 40.31 | 0.97 | 1.02 | 22.05 | 44.575 |
| 1542 百万 | 8.63 | 63.24 | 93.30 | 89.05 | 18.34 | 35.76 | 0.93 | 0.98 | 17.48 | 42.16 |

表 3. 多个数据集上的零样本结果。这些结果均未进行训练或微调。PTB 和 WikiText-2 结果来自 (Gong et al., 2018)。CBT 结果来自 (Bajgar et al., 2016)。LAMBADA 准确率结果来自 (Hoang et al., 2018)，LAMBADA 困惑度结果来自 (Grave et al., 2016)。其他结果来自 (Dai et al., 2019)。

这在 WebText 中极为罕见——在 400 亿字节中仅出现 26 次。我们在表 3 中报告了使用可逆去分词器的主要结果，这些去分词器尽可能多地消除了此类分词/预处理伪像。由于这些去分词器是可逆的，我们仍然可以计算数据集的对数概率，并且它们可以被视为一种简单的领域自适应形式。我们观察到，使用这些去分词器后，GPT-2 的困惑度提升了 2.5 到 5。

WebText 语言模型在跨领域和数据集方面表现出色，在零样本设置下，在 8 个数据集集中的 7 个上提升了最佳性能。在只有 100 万到 200 万个训练标记的小型数据集（例如 Penn Treebank 和 WikiText-2）上也能看到显著的提升。在用于测量长期依赖关系的数据集（例如 LAMBADA (Paperno 等人，2016 年) 和儿童图书测试 (Hill 等人，2015 年)）上也能看到显著的提升。我们的模型在十亿词基准测试 (Chelba 等人，2013 年) 上的表现仍然明显逊于先前的研究。这可能是因为它既是最大的数据集，又具有一些最具破坏性的预处理——1BW 的句子级别改组会删除所有长程结构。

儿童图书测试 (CBT) (Hill 等人，2015) 旨在检验语言模型 (LM) 对不同类别词汇的表现：命名实体、名词、动词和介词。CBT 并不将困惑度作为评估指标，而是报告自动构建的完形填空测试的准确率。完形填空测试的任务是预测省略单词的 10 个可能选项中哪一个是正确的。遵循原始论文中介绍的 LM 方法，我们根据 LM 计算每个选项的概率以及基于该选项的句子其余部分的概率，并预测概率最高的选项。如图 2 所示，随着模型规模的增加，性能稳步提升，并缩小了与人类在该测试中的表现之间的大部分差距。数据重叠分析显示，CBT 测试集书籍之一，拉迪亚德·吉卜林的《丛林之书》，是在 WebText 中，因此我们报告验证集的结果，因为两者没有显著重叠。GPT-2 在普通名词识别上取得了 93.3% 的准确率，在命名实体识别上取得了 89.1% 的准确率，并达到了新的 SOTA 最高水平。我们使用去标记器来移除 CBT 中 PTB 风格的标记伪像。

3.2. 儿童读物测试

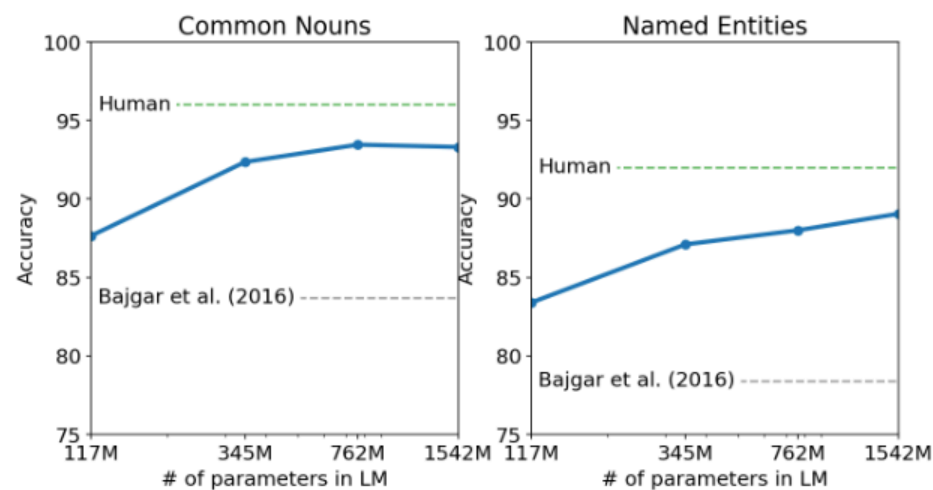


图 2. 儿童图书测试中模型容量与性能的关系。人类表现数据来自 Bajgar 等人 (2016 年)，而非原始论文中低得多的估计值。

3.3. 兰巴达

LAMBADA 数据集 (Paperno 等人，2016) 测试系统对文本中长距离依赖关系进行建模的能力。该任务是预测句子的最后一个单词，人类至少需要 50 个上下文标记才能成功预测。GPT-2 将最佳结果的困惑度 (perplexity) 从 99.8 (Grave 等人，2016) 提升至 8.6，并将语言模型 (LM) 在此测试中的准确率从 19% (Dehghani 等人，2018) 提升至 52.66%。调查 GPT-2 的错误表明，大多数预测是句子的有效延续，但不是有效的最后一个单词。这表明语言模型 (LM) 没有使用额外的有用约束，即单词必须是句子的最后一个单词。添加停用词过滤器作为近似值，可以进一步将准确率提高到 63.24%，使该任务的整体最佳结果提高 4%。先前的最佳实践 (Hoang 等人，2018) 使用了不同的受限预测设置，其中模型的输出仅限于上下文中出现的单词。对于 GPT-2 来说，这种限制有害无益。

因为 19% 的答案与上下文无关。我们使用未经预处理的数据集版本。

3.4. 维诺格勒模式挑战

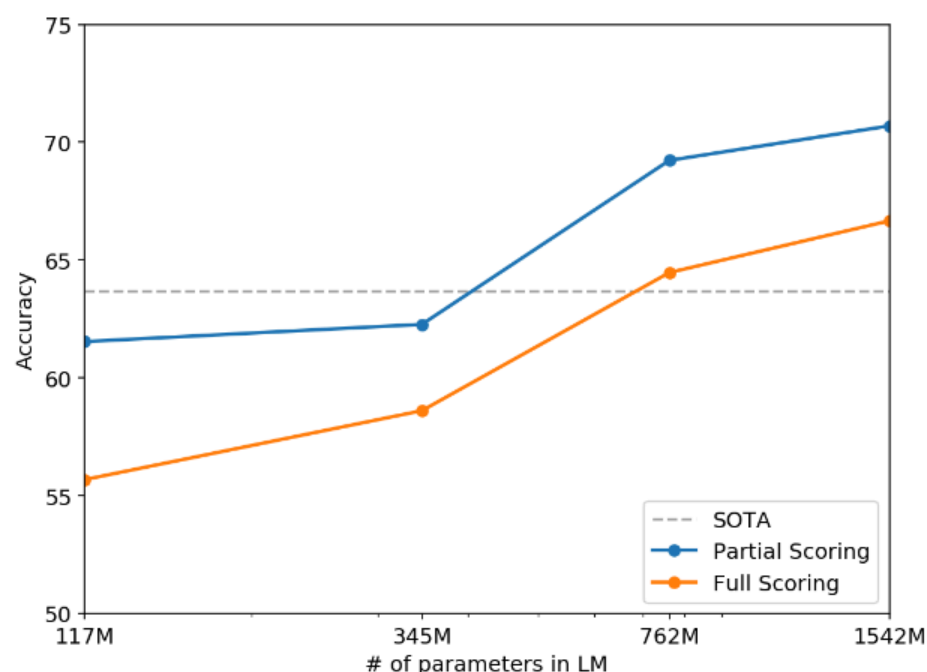


图 3. Winograd Schema Challenge 上的表现与模型容量的关系。

Winograd Schema 挑战赛 (Levesque 等人, 2012) 旨在通过衡量系统解决文本歧义的能力来衡量其进行常识推理的能力。最近, Trinh & Le (2018) 利用语言模型 (LM) 在该挑战赛中取得了显著进展, 他们以更高的概率预测了歧义的解决。我们遵循他们的问题表述, 并在图 3 中可视化了我们模型分别使用完整评分和部分评分技术的性能。GPT-2 将最佳准确率提高了 7%, 达到 70.70%。由于数据集规模较小, 只有 273 个示例, 因此我们建议您阅读 Trichelair 等人 (2018) 的文章, 以帮助理解这一结果。

3.5. 阅读理解

Reddy 等人 (2018) 的对话问答数据集 (CoQA) 包含来自 7 个不同领域的文档, 并配有提问者 and 回答者之间关于该文档的自然语言对话。CoQA 测试模型的阅读理解能力, 以及基于对话历史 (例如 “为什么?”) 回答问题的能力。

GPT-2 的贪婪解码以文档、相关对话历史记录和最终标记 A: 为条件, 在开发集上达到了 55 F1 的准确率。这相当于甚至超过了 4 个基线系统中的 3 个, 而无需使用 127,000 多个手动收集的问答对来训练这些基线系统。监督学习的 SOTA, 一个基于 BERT 的系统 (Devlin 等人,

| | R-1 | R-2 | RL | R-平均值 | |
|---------|--------------|-------|----------------|-----------|---------------|
| 自下而上总和 | 41.22 | 18.68 | 38.34 | 32.75 | Lede-3 40.38 |
| 17.66 | 36.62 | 31.55 | Seq2Seq + Attn | 31.33 | 11.81 28.83 |
| 23.99 | GPT-2 TL;DR: | 29.34 | 8.27 | 26.58 | 21.40 Random- |
| 3 28.78 | 8.63 | 25.52 | 20.98 | GPT-2 无提示 | 21.58 4.03 |
| 19.47 | 15.03 | | | | |

表 4. 使用 ROUGE F1 指标在 CNN 和每日邮报数据集上测量的摘要性能。Bottom-Up Sum 是 Gehrmann 等人, 2018 年提出的 SOTA 模型。

2018 年, GPT-2 的 F1 成绩已接近人类的 89 分。虽然对于一个无需任何监督训练的系统来说, GPT-2 的表现令人兴奋, 但对其答案和错误的一些检查表明, GPT-2 通常使用基于简单检索的启发式方法, 例如在回答 “谁” 的问题时, 使用文档中的名称作为答案。

3.6. 总结

我们在 CNN 和每日邮报数据集 (Nallapati et al., 2016) 上测试了 GPT-2 执行摘要的能力。为了诱导摘要行为, 我们在文章后添加文本 TL;DR:, 并使用 Top-k 随机抽样 (Fan et al., 2018) 生成 100 个标记, 其中 $k = 2$, 这可以减少重复并鼓励比贪婪解码更抽象的摘要。我们使用这 100 个标记中生成的前 3 个句子作为摘要。虽然从质量上讲, 生成的摘要与摘要类似 (如表 14 所示), 但它们通常关注文章中的最新内容或混淆具体细节, 例如有多少辆汽车卷入了车祸, 或者标志是在帽子还是衬衫上。在常见的 ROUGE 1,2,L 指标上, 生成的摘要才刚刚开始接近经典神经基线的性能, 并且仅仅略胜于从文章中选择 3 个随机句子。当删除任务提示时, GPT-2 的性能在总体指标上下降了 6.4 分, 这表明它能够使用自然语言在语言模型中调用特定任务的行为。

3.7. 翻译

我们测试 GPT-2 是否已经开始学习如何从一种语言翻译成另一种语言。为了帮助它推断这是所需的任务, 我们以英语格式的示例对作为条件, 对语言模型进行条件化。

句子 = 法语句子, 然后是 fi-
最终的英语句子提示 = 我们通过贪婪解码从模型中采样, 并使用第一个生成的句子作为翻译。在 WMT-14 英法测试集上, GPT-2 的 BLEU 为 5, 这比之前在无监督词汇翻译研究中推断出的双语词典逐字替换的结果略差。

| 问题 | 生成答案正确概率 |
|---|----------|
| 谁写了《物种起源》这本书？ 查尔斯·达尔文 3 83.4% 谁是 ubuntu 项目的创始人？ 马克·沙特尔沃思 3 82.0% 谁是绿湾包装工队的四分卫？ 阿隆·罗杰斯 3 81.1% 熊猫是哪个国家的国兽？ 中国 3 76.8% 谁提出了相对论？ 阿尔伯特·爱因斯坦 3 76.4% 第一部星球大战电影是什么时候上映的？ 1977 3 71.4% 瑞典最常见的血型是什么？ A 7 70.6% 谁被认为是精神分析的创始人？ 西格蒙德·弗洛伊德 3 69.3% 谁在 1969 年迈出了登上月球的第一步？ 尼尔·阿姆斯特朗 3 66.8% 谁是英国最大的连锁超市？ 乐购 3 65.3% shalom 在英语中的意思是什么？ 和平 3 64.0% 谁是《孙子兵法》的作者？ 孙子 3 59.6% 美国陆地面积最大的州是哪个？ 加利福尼亚州 7 59.2% 绿藻是哪种繁殖类型的一个例子？ 孤雌生殖 7 56.5% 维克拉姆·萨姆瓦特日历在哪个国家是官方的？ 印度 3 55.6% 谁主要负责撰写《独立宣言》？ 托马斯·杰斐逊 3 53.3% 哪个州构成了蒙大拿州的西部边界？ 蒙大拿州 7 52.3% 谁在《权力的游戏》中扮演戴佛斯爵士？ 彼特·丁拉基 7 52.1% 谁任命了联邦储备系统主席？ 珍妮特·耶伦 7 51.5% 说明将一个细胞核分成两个基因相同的细胞核的过程？ 有丝分裂 3 50.7% 谁获得了 NBA 最多的 MVP 奖？ 迈克尔·乔丹 7 50.2% 哪条河与罗马城有关？ 台伯河 3 48.6% 谁是第一个被弹劾的总统？ 安德鲁·约翰逊 3 48.3% 谁是 2017 年国土安全部的负责人？ 约翰·凯利 3 47.0% 欧盟的共同货币名称是什么？ 欧元 3 46.8% 星球大战中的皇帝叫什么名字？ 帕尔帕廷 3 46.5% 去靶场射击需要持枪许可证吗？ 否 3 46.4% 谁在 1859 年提出了进化论作为生物发展的基础？ 查尔斯·达尔文 3 45.7% 俄罗斯爆炸的核电站？ 切尔诺贝利 3 45.7% 谁在《终结者》第一部中扮演约翰·康纳？ 阿诺德·施瓦辛格 7 45.2% | |

表 5. GPT-2 在自然问题开发集上生成的 30 个最有信心的答案，按 GPT-2 给出的概率排序。根据第 4 节中描述的程序，这些问题均未出现在 WebText 中。

(Conneau 等人, 2017b)。在 WMT-14 法英测试集上，GPT-2 能够利用其非常强大的英语语言模型表现得更好，达到 11.5 BLEU。这优于 (Artetxe 等人, 2017) 和 (Lample 等人, 2017) 中的几个无监督机器翻译基线，但仍然比当前最佳无监督机器翻译方法 (Artetxe 等人, 2019) 的 33.5 BLEU 差很多。这项任务的表现令我们感到惊讶，因为我们特意从 WebText 中删除了非英语网页作为过滤步骤。为了证实这一点，我们在 WebText 上运行了一个字节级语言检测器，它只检测到 10MB 的法语数据，这比以前的无监督机器翻译研究中常见的单语法语语料库小大约 500 倍。

2019 年发表的论文《GPT-2 的语义理解》(2019 年) 是一个很有前景的资源，可以更定量地测试这一点。与翻译类似，语言模型的上下文以示例问答对为种子，这有助于模型推断数据集的简短回答风格。使用 SQUAD 等阅读理解数据集中常用的精确匹配指标进行评估时，GPT-2 的正确率仅为 4.1%。作为比较，最小的模型的准确率不超过一个极其简单的基线模型的 1.0%，该基线模型会返回每种问题类型（谁、什么、哪里等）的最常见答案。GPT-2 的正确率是后者的 5.3 倍，这表明模型容量是迄今为止神经系统在此类任务中表现不佳的主要因素。GPT-2 为其生成的答案分配的概率经过了良好的校准，并且 GPT-2 对其最有信心的 1% 问题的准确率为 63.1%。表 5 显示了 GPT-2 在开发集问题上生成的 30 个最有信心的答案。GPT-2 的性能仍然远远低于 30% 到 50% 的开放领域问答系统，这些系统将信息检索与提取文档问答相结合 (Alberti 等人, 2019 年)。

3.8. 问答

测试语言模型中包含哪些信息的一种潜在方法是评估它对事实类问题生成正确答案的频率。之前在神经系统中展示的这种行为，所有信息都存储在参数中，例如“神经对话模型”(Vinyals & Le, 2015)，由于缺乏高质量的评估数据集，报告的结果只是定性的。最近推出的“自然问题”数据集 (Kwiatkowski 等人,

Alec³ 之前以为自己擅长随机问答，但在与 GPT-2 相同的环境下进行测试时，他正确回答了 100 个随机样本中的 17 个。实际上，他只答对了 14 个，但他

³应该得到另外 3 个

²<https://github.com/CLD20wners/cld2>

| | PTB | WikiText-2 | enwik8 | text8 | Wikitext-103 | 1BW |
|------------|-------|------------|--------|-------|--------------|--------|
| 数据集训练 | 2.67% | 0.66% | 7.50% | 2.34% | 9.09% | 13.19% |
| WebText 训练 | 0.88% | 1.63% | 6.31% | 3.94% | 2.42% | 3.75% |

表 6.测试集 8 克与训练集重叠的百分比。

4. 泛化与记忆

计算机视觉领域的最新研究表明，常见的图像数据集包含大量近似重复的图像。例如，CIFAR-10 的训练集和测试集图像重叠度高达 3.3% (Barz & Denzler, 2019)。这导致机器学习系统泛化性能被夸大。随着数据集规模的扩大，这个问题出现的可能性越来越大，这表明 WebText 也可能出现类似的现象。因此，分析有多少测试数据也出现在训练集数据中至关重要。

为了研究这一点，我们创建了包含 8 个 WebText 训练集标记的布隆过滤器。为了提高召回率，我们将字符串规范化为仅包含小写字母数字单词，并以单个空格作为分隔符。布隆过滤器的构造使得误报率的上限为 $\frac{1}{2^8}$ 。我们通过生成 100 万个字符串进一步验证了其较低的误报率，其中过滤器未发现任何字符串。

利用这些布隆过滤器，我们可以计算出给定数据集中 8 元语法在 WebText 训练集中出现的百分比。表 6 展示了常见语言模型 (LM) 基准测试集的重叠度分析。常见语言模型 (LM) 数据集的测试集与 WebText 训练集的重叠度在 1% 到 6% 之间，平均重叠度为 3.2%。令人惊讶的是，许多数据集与其自身的训练样本的重叠度更大，平均重叠度为 5.9%。

我们的方法针对召回率进行了优化，虽然手动检查重叠部分会发现许多常用短语，但也存在许多由于数据重复而导致的较长匹配。这并非 WebText 独有。例如，我们发现 WikiText-103 的测试集中有一篇文章也包含在训练数据集中。由于测试集中只有 60 篇文章，因此重叠率至少为 1.6%。更令人担忧的是，根据我们的程序，1BW 与其自身的训练集重叠率接近 13.2%。

在 Winograd Schema 挑战赛中，我们仅发现 10 个模式与 WebText 训练集存在 8-gram 重叠。其中 2 个是虚假匹配。在剩下的 8 个模式中，只有 1 个出现在任何上下文中，

⁴很大一部分额外重叠是由于编辑在多篇文章中重复使用了一些具有共同主题的段落，例如朝鲜战争中的各个战役。

透露了答案。

对于 CoQA，新闻领域中约 15% 的文档已在 WebText 中，模型在这些文档上的表现 F1 值提高了约 3。CoQA 的开发集指标报告了 5 个不同领域的平均性能，由于不同领域之间存在重叠，我们测得 F1 值提升了约 0.5-1.0。但是，由于 CoQA 是在 WebText 链接截止日期之后发布的，因此 WebText 中没有实际的训练问题或答案。

在 LAMBADA 上，平均重叠率为 1.2%。GPT-2 在重叠率超过 15% 的样本上的表现大约提高了 2 个困惑度。如果在排除所有重叠率发生变化的样本后重新计算指标，困惑度将从 8.6 降至 8.7，准确率也将从 63.2% 降至 62.9%。整体结果出现如此微小的变化，很可能是因为每 200 个样本中只有 1 个具有显著的重叠度。

总体而言，我们的分析表明，WebText 训练数据与特定评估数据集之间的数据重叠会为报告结果带来微小但持续的益处。然而，对于大多数数据集，我们并未发现其重叠程度明显高于标准训练集和测试集之间现有的重叠程度，如表 6 所示。

理解并量化高度相似的文本如何影响性能是一个重要的研究问题。更强大的去重技术（例如可扩展模糊匹配）也有助于更好地解答这些问题。目前，我们建议在为新的 NLP 数据集创建训练和测试样本时，使用基于 n-gram 重叠的去重技术作为重要的验证步骤和健全性检查。

确定 WebText 语言模型的性能是否归因于记忆的另一种潜在方法是检查它们在自身保留集上的性能。如图 4 所示，WebText 在训练集和测试集上的性能相似，并且随着模型规模的增加而共同提升。这表明即使是 GPT-2 在 WebText 上仍然在很多方面存在欠拟合的情况。

GPT-2 还能撰写关于发现会说话的独角兽的新闻报道。表 13 提供了一个示例。

5.相关工作

这项研究的很大一部分内容是测量在更大的数据集上训练的大型语言模型的性能。

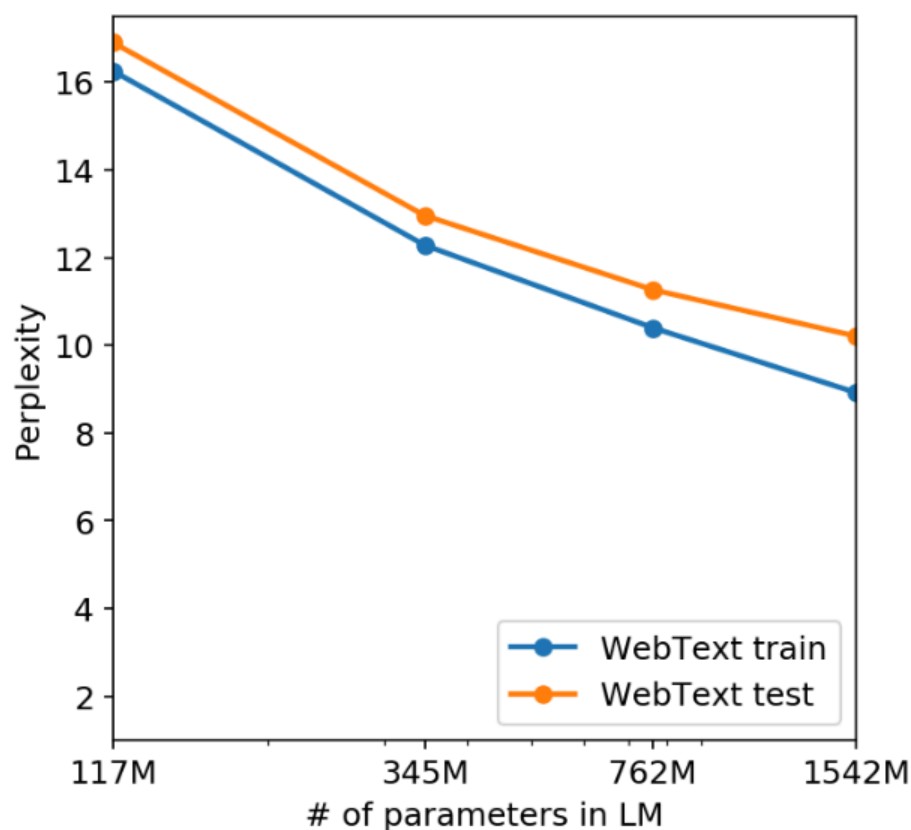


图 4. 在 WebText 上训练的 LM 的性能与模型大小的关系。

与 Jozefowicz 等人 (2016 年) 的工作类似, 他们在 10 亿词基准上扩展了基于 RNN 的语言模型。Bajgar 等人

(2016 年) 此前也曾利用古腾堡计划创建了一个更大的训练数据集来补充标准训练数据集, 从而提升了儿童读物测试的结果。Hestness 等人 (2017 年) 对各种深度学习模型的性能如何随着模型容量和数据集大小的变化进行了深入分析。我们的实验虽然在不同任务中噪声更大, 但表明目标函数的子任务也呈现出类似的趋势, 并持续到 10 亿以上的参数范围。

生成模型中一些有趣的学习功能此前已有文献记载, 例如 Karpathy 等人 (2015) 提出的 RNN 语言模型中的单元执行行宽追踪和引用/评论检测。刘等人 (2018) 的观察结果对我们的工作更具启发: 一个经过训练生成维基百科文章的模型也学会了在不同语言之间翻译人名。

之前的研究探索了过滤和构建大型网页文本语料库的替代方法, 例如 iWeb 语料库 (Davies, 2018)。

语言任务的预训练方法已有大量研究。除了引言中提到的方法外, GloVe (Pennington 等人, 2014) 将词向量表征学习扩展到所有 Common Crawl 数据集。一项早期关于文本深度表征学习的有影响力的研究是 Skip-thought Vectors (Kiros 等人, 2015)。McCann 等人

(2017) 探索了源自机器翻译模型的表征的使用, Howard & Ruder (2018) 改进了 (Dai & Le, 2015) 基于循环神经网络 (RNN) 的微调方法。(Conneau et al., 2017a) 研究了自然语言推理模型学习到的表征的迁移性能,

(Subramanian et al., 2018) 探索了大规模多任务训练。(Ramachandran 等人, 2016) 证明, 使用预训练的语言模型作为编码器和解码器进行初始化, 对 seq2seq 模型大有裨益。近期研究表明, 语言模型预训练在针对诸如聊天对话和基于对话的问答系统等高难度生成任务进行微调时, 也同样有效 (Wolf 等人, 2019) (Dinan 等人, 2018)。

6. 讨论

大量研究致力于学习 (Hill 等人, 2016)、理解 (Levy & Goldberg, 2014) 以及批判性地评估 (Wieting & Kiela, 2019) 监督和无监督预训练方法的表征。我们的结果表明, 无监督任务学习是另一个有前景的研究领域。这些发现可能有助于解释预训练技术在下游 NLP 任务中取得的广泛成功, 因为我们表明, 在极限情况下, 其中一种预训练技术无需监督调整或修改即可直接开始学习执行任务。

在阅读理解方面, GPT-2 的性能在零样本环境下与监督基线模型相媲美。然而, 在其他任务 (例如摘要) 上, 虽然它在定性上能够完成任务, 但根据定量指标来看, 其性能仍然处于初级阶段。虽然作为一项研究成果具有启发性, 但就实际应用而言, GPT-2 的零样本性能仍然远远达不到实用水平。我们已经研究了 WebText 语言模型在许多经典 NLP 任务上的零样本性能, 但还有许多其他任务可以评估。毫无疑问, 在许多实际任务中, GPT-2 的性能仍然不比随机任务好。即使在我们评估的常见任务 (例如问答和翻译) 中, 语言模型也只有在具备足够容量时才能超越普通的基线模型。

虽然零样本性能为 GPT-2 在许多任务上的潜在性能设定了基准, 但微调的上限尚不清楚。在某些任务上, GPT-2 的完全抽象输出与基于提取指针网络 (Vinyals 等人, 2015) 的输出存在显著差异, 而后者目前在许多问答和阅读理解数据集上都处于领先地位。鉴于 GPT 微调的先前成功, 我们计划在 decaNLP 和 GLUE 等基准测试上研究微调, 尤其是在目前尚不清楚额外的

GPT-2 的训练数据和容量足以克服 BERT (Devlin et al., 2018) 所展示的单向表示的低效率。

7. 结论

当大型语言模型在足够庞大且多样化的数据集上进行训练时，它能够在众多领域和数据集上表现良好。GPT-2 在 8 个测试语言建模数据集中的 7 个上，零样本训练达到了最佳性能。该模型在零样本训练中能够执行的任务多样性表明，经过训练以最大化足够多样化的文本语料库的似然性，高容量模型开始学习如何在无需明确监督的情况下执行数量惊人的任务。

致谢

感谢所有撰写本文、分享链接并在 WebText 中点赞的朋友们。数百万人参与了 GPT-2 训练数据的创建。同时，也感谢所有帮助我们构建训练基础设施的 Google 员工，包括 Zak Stone、JS Riehl、Jonathan Hseu、Russell Power、Yulong Cheng、Noam Shazeer、Solomon Boulos、Michael Banfield、Aman Gupta、Daniel Sohn 等等。最后，感谢为本文草稿提供反馈的各位：Jacob Steinhardt、Sam Bowman、Geoffrey Irving 和 Madison May。

参考

Al-Rfou, R., Choe, D., Constant, N., Guo, M. 和 Jones, L. 具有更深层次自我关注的字符级语言建模。arXiv 预印本 arXiv:1808.04444, 2018。

Alberti, C., Lee, K. 和 Collins, M. 自然问题的 bert 基线。arXiv 预印本 arXiv:1901.08634, 2019 年。

Alcorn, MA, Li, Q., Gong, Z., Wang, C., Mai, L., Ku, W.-S. 和 Nguyen, A. Strike (with) a pose: 神经网络很容易被熟悉物体的奇怪姿势所欺骗。arXiv 预印本 arXiv:1811.11553, 2018 年。

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G. 等。深度语音 2: 端到端语音识别

英语和普通话。在国际机器学习会议上，第 173-182 页，2016 年。

Artetxe, M., Labaka, G., Agirre, E. 和 Cho, K. 无监督神经机器翻译。arXiv 预印本 arXiv:1710.11041, 2017 年。

Artetxe, M., Labaka, G. 和 Agirre, E. 一种有效的无监督机器翻译方法。arXiv 预印本 arXiv:1902.01313, 2019 年。

下载和使用小模型的初步代码可在 <https://github.com/openai/gpt-2> 获取。

Ba, JL, Kiros, JR 和 Hinton, GE 层标准化。arXiv 预印本 arXiv:1607.06450, 2016。

Bajgar, O., Kadlec, R. 和 Kleindienst, J. 拥抱数据丰富：用于阅读理解的 Booktest 数据集。arXiv 预印本 arXiv:1610.00956, 2016 年。

Barz, B. 和 Denzler, J. 我们是否使用测试数据进行训练？清除 cifar 中的近似重复项。arXiv 预印本 arXiv:1902.00423, 2019 年。

Bengio, Y., Ducharme, R., Vincent, P. 和 Jauvin, C. 一种神经概率语言模型。《机器学习研究杂志》，2003 年 2 月 3 日：1137–1155 页。

Bowman, SR, Pavlick, E., Grave, E., Van Durme, B., Wang, A., Hula, J., Xia, P., Pappagari, R., McCoy, RT, Patel, R. 等。

寻找 elmo 的朋友：超越语言建模的句子级预训练。arXiv 预印本 arXiv:1812.10860, 2018 年。

Caruana, R. 多任务学习。机器学习，28(1):41–75, 1997。

Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P. 和 Robinson, T. 用于衡量统计语言建模进展的十亿字基准。arXiv 预印本 arXiv:1312.3005, 2013 年。

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K. 和 Kuksa, P. 自然语言处理（几乎）来自

scratch。《机器学习研究杂志》，2011 年 8 月 12 日：2493–2537。

Conneau, A., Kiela, D., Schwenk, H., Barrault, L. 和 Bordes, A. 从自然语言推理数据中监督学习通用句子表示。arXiv 预印本 arXiv:1705.02364, 2017a。

Conneau, A., Lample, G., Ranzato, M., Denoyer, L. 和 Jégou, H. 没有平行数据的词语翻译。arXiv 预印本 arXiv:1710.04087, 2017b。

Dai, AM 和 Le, QV 半监督序列学习。在

神经信息处理系统的进展，第 3079-3087 页，2015 年。

Dai, Z., Yang, Z., Yang, Y., Cohen, WW, Carbonell, J., Le, QV 和 Salakhutdinov, R. Transformer-xl: 超越固定长度上下文的专注语言模型。arXiv 预印本 arXiv:1901.02860, 2019 年。

Davies, M. 140 亿字的 iweb 语料库。

<https://corpus.byu.edu/iWeb/>, 2018 年。

Dehghani, M., Gouws, S., Vinyals, O., Uszkoreit, J. 和 Kaiser, Ł. 万能变压器。arXiv 预印本 arXiv:1807.03819, 2018。

Devlin, J., Chang, M.-W., Lee, K. 和 Toutanova, K. Bert: 用于语言理解的深度双向转换器的预训练。arXiv 预印本 arXiv:1810.04805, 2018 年。

Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M. 和 Weston, J. 维基百科奇才：知识驱动的对话代理。arXiv 预印本 arXiv:1811.01241, 2018 年。

Fan, A., Lewis, M. 和 Dauphin, Y. 分层神经故事生成。arXiv 预印本 arXiv:1805.04833, 2018 年。

Finn, C., Abbeel, P. 和 Levine, S. 模型无关元学习, 用于快速适应深度网络。arXiv 预印本 arXiv:1703.03400, 2017 年。

Gehrmann, S., Deng, Y. 和 Rush, AM 自下而上的抽象概括。arXiv 预印本 arXiv:1808.10792, 2018 年。

Gillick, D., Brunk, C., Vinyals, O. 和 Subramanya, A. 从字节进行多语言处理。arXiv 预印本 arXiv:1512.00103, 2015 年。

Gong, C., He, D., Tan, X., Qin, T., Wang, L. 和 Liu, T.-Y. Frage: 频率不可知的词语表征。载于《神经信息处理系统进展》, 第 1341-1352 页, 2018 年。

Grave, E., Joulin, A. 和 Usunier, N. 使用连续缓存改进神经语言模型。arXiv 预印本 arXiv:1612.04426, 2016 年。

He, K., Zhang, X., Ren, S., 和 Sun, J. 深度身份映射

残差网络。欧洲计算机视觉会议, 第 630-645 页。Springer, 2016 年。

Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M., Ali, M., Yang, Y. 和 Zhou, Y. 根据经验, 深度学习缩放是可预测的。arXiv 预印本 arXiv:1712.00409, 2017。

Hill, F., Bordes, A., Chopra, S. 和 Weston, J. 金发姑娘原则: 用外显记忆表征阅读儿童读物。arXiv 预印本 arXiv:1511.02301, 2015 年。

Hill, F., Cho, K., 和 Korhonen, A. 从无标签数据中学习句子的分布式表示。arXiv 预印本 arXiv:1602.03483, 2016 年。

Hoang, L., Wiseman, S. 和 Rush, AM 实体跟踪提高了完形填空式阅读理解能力。arXiv 预印本 arXiv:1810.02891, 2018 年。

Howard, J. 和 Ruder, S. 通用语言模型微调

文本分类。载于《计算语言学协会第 56 届年会论文集 (第 1 卷: 长篇论文)》, 第 1 卷, 第 328-339 页, 2018 年。

Jelinek, F. 和 Mercer, RL, 《从稀疏数据中插值估计马尔可夫源参数》。载于《实践模式识别研讨会, 荷兰阿姆斯特丹: 北荷兰, 1980 年 5 月。

Jia, R. 和 Liang, P. 评估阅读的对抗性例子理解系统。arXiv 预印本 arXiv:1707.07328, 2017 年。

Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N. 和 Wu, Y. 探索语言建模的极限。arXiv 预印本 arXiv:1602.02410, 2016 年。

Kaiser, L., Gomez, AN, Shazeer, N., Vaswani, A., Parmar, N., Jones, L. 和 Uszkoreit, J. 一种可以学习所有这些的模型。arXiv 预印本 arXiv:1706.05137, 2017。

Karpathy, A., Johnson, J. 和 Fei-Fei, L. 可视化和理解

常态循环网络。arXiv 预印本 arXiv:1506.02078, 2015。

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, AA, Milan, K., Quan, J., Ramalho, T., GrabskaBarwinska, A. 等。克服灾难性遗忘神经网络。《美国国家科学院院刊》, 第 201611835 页, 2017 年。

Kiros, R., Zhu, Y., Salakhutdinov, RR, Zemel, R., Urtasun, R., Torralba, A. 和 Fidler, S. 跳过思想向量。神经信息处理系统进展, 第 3294-3302 页, 2015 年。

Krizhevsky, A., Sutskever, I. 和 Hinton, 基于深度卷积神经网络的 GE Imagenet 分类。载于《神经信息处理系统进展》, 第 1097-1105 页, 2012 年。

Kwiatkowski, T., Palomaki, J., Rhinehart, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Kelcey, M., Devlin, J. 等人。《自然问题: 问答研究的基准》。2019 年。Lake, BM, Ullman, TD, Tenenbaum, JB, 和 Gershman, SJ

构建像人类一样学习和思考的机器。《行为与脑科学》, 第 40 卷, 2017 年。

Lample, G., Conneau, A., Denoyer, L. 和 Ranzato, M. 仅使用单语语料库进行无监督机器翻译。

arXiv 预印本 arXiv:1711.00043, 2017。

Levesque, H., Davis, E., 和 Morgenstern, L. winograd 模式挑战。在 2012 年第十三届知识表示与推理原则国际会议上。

Levy, O. 和 Goldberg, Y. 神经词嵌入作为隐式模型 trix 因式分解。载于《神经信息处理系统进展》, 第 2177-2185 页, 2014 年。

Liu, PJ, Saleh, M., Pot, E., Goodrich, B., Sepassi, R., Kaiser, L. 和 Shazeer, N. 通过汇总长序列生成维基百科。arXiv 预印本 arXiv:1801.10198, 2018 年。

McCann, B., Bradbury, J., Xiong, C. 和 Socher, R. 翻译中的学习: 语境化词向量。载于《神经信息处理系统进展》, 第 6294-6305 页, 2017 年。

McCann, B., Keskar, NS, Xiong, C. 和 Socher, R. 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。

Merity, S., Xiong, C., Bradbury, J. 和 Socher, R. 指针哨兵混合模型。arXiv 预印本 arXiv:1609.07843, 2016 年。

Mikolov, T., Sutskever, I., Chen, K., Corrado, GS 和 Dean, J. 单词和短语的分布式表示及其

组合性。载于《神经信息处理系统进展》, 第 3111-3119 页, 2013 年。

Nallapati, R., Zhou, B., Gulcehre, C., Xiang, B. 等人。基于序列到序列 RNN 及其他算法的抽象文本摘要。arXiv 预印本 arXiv:1602.06023, 2016 年。

Paperno, D., Kruszewski, G., Lazaridou, A., Pham, QN, Bernardi, R., Pezzelle, S., Baroni, M., Boleda, G. 和 Fern´andez, R. Lambada 数据集: 需要广泛话语背景的词汇预测。arXiv 预印本 arXiv:1606.06031, 2016 年。

Pennington, J., Socher, R. 和 Manning, C. Glove: 全球载体

用于词汇表征。载于 2014 年自然语言处理经验方法会议 (EMNLP) 论文集, 第 1532-1543 页, 2014 年。

Peters, ME 和 Lecocq, D. 使用不同的特征进行内容提取集。载于《第 22 届万维网国际会议论文集》，第 89-90 页。ACM, 2013 年。

Peters, ME, Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. 和 Zettlemoyer, L. 深度语境化词汇表征。arXiv 预印本 arXiv:1802.05365, 2018 年。

Radford, A., Jozefowicz, R. 和 Sutskever, I. 学习生成评论和发现情绪。arXiv 预印本 arXiv:1704.01444, 2017 年。

Radford, A., Narasimhan, K., Salimans, T. 和 Sutskever, I. 通过生成预训练提高语言理解能力。2018.

Ramachandran, P., Liu, PJ 和 Le, QV, 用于序列到序列学习的无监督预训练。arXiv 预印本 arXiv:1611.02683, 2016 年。

Recht, B., Roelofs, R., Schmidt, L. 和 Shankar, V. cifar-10 分类器能推广到 cifar-10 吗? arXiv 预印本 arXiv:1806.00451, 2018 年。

Reddy, S., Chen, D. 和 Manning, CD Coqa: 对话问答挑战。arXiv 预印本 arXiv:1808.07042, 2018 年。

Schwartz, R., Sap, M., Konstas, I., Zilles, L., Choi, Y., 和 Smith, NA Story 完形填空任务: Uw nlp 系统。在第二届词汇、句子和篇章语义学链接模型研讨会, 第 52-55 页, 2017 年。

参见 A., Liu, PJ 和 Manning, CD 直奔主题: 使用指针生成网络进行总结。arXiv 预印本 arXiv:1704.04368, 2017 年。

Sennrich, R., Haddow, B. 和 Birch, A. 使用子词单元的稀有词神经机器翻译。arXiv 预印本 arXiv:1508.07909, 2015 年。

Subramanian, S., Trischler, A., Bengio, Y. 和 Pal, CJ 通过大规模学习通用分布式句子表征
大规模多任务学习。arXiv 预印本 arXiv:1804.00079, 2018 年。

Sutskever, I., Vinyals, O. 和 Le, QV, 《基于神经网络的序列到序列学习》。载于《神经信息处理系统进展》, 第 3104-3112 页, 2014 年。

Sutskever, I., Jozefowicz, R., Gregor, K., Rezende, D., Lillicrap, T. 和 Vinyals, O. 走向有原则的无监督学习。
arXiv 预印本 arXiv:1511.06440, 2015。

Trichelair, P., Emami, A., Cheung, JCK, Trischler, A., Suleman, K. 和 Diaz, F. 关于自然语言理解中常识推理的评估。arXiv 预印本 arXiv:1811.01778, 2018 年。

Trinh, TH 和 Le, QV, 一种简单的常识推理方法。arXiv 预印本 arXiv:1806.02847, 2018 年。

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, AN, Kaiser, Ł. 和 Polosukhin, I. 注意是你所需要的一切。《神经信息处理系统进展》, 第 5998-6008 页, 2017 年。

Vinyals, O. 和 Le, Q. 一种神经对话模型。arXiv 预印本 arXiv:1506.05869, 2015 年。

Vinyals, O., Fortunato, M. 和 Jaitly, N. 指针网络。在神经信息处理系统进展, 第 2692-2700 页, 2015 年。

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O. 和 Bowman, SR Glue: 用于自然语言理解的多任务基准和分析平台。arXiv 预印本 arXiv:1804.07461, 2018 年。

Weston, JE. 基于对话的语言学习。载于《神经信息处理系统进展》, 第 829-837 页, 2016 年。

Wieting, J. 和 Kiela, D. 无需训练: 探索用于句子分类的随机编码器。arXiv 预印本 arXiv:1901.10444, 2019 年。

Wolf, T., Sanh, V., Chaumond, J. 和 Delangue, C. Transfertransfo: 一种基于神经网络对话代理的迁移学习方法。arXiv 预印本 arXiv:1901.08149, 2019 年。

Yogatama, D., d'Áutume, C. d. M., Connor, J., Kocisky, T., Chrzanowski, M., Kong, L., Lazaridou, A., Ling, W., Yu, L., Dyer, C. 等。学习和评估一般语言智能。arXiv 预印本 arXiv:1901.11373, 2019。

8. 附录 A：示例

8.1. 模型容量

为了补充图 4 中显示的 WebText 上较大的 LM 所报告的困惑度增益，表 7 至表 11 并排展示了最小的 WebText LM 和 GPT-2 在随机未见过的 WebText 测试集文章上的完成情况。

8.2. 文本记忆

我们观察到 GPT-2 对数据集中重复多次的长字符串（例如名人名言或演讲）表现出一些记忆行为。例如，当以葛底斯堡演说的前半句（在 WebText 中出现约 40 次）为条件时，GPT-2 的 argmax 解码可以恢复语音。即使在不截断的情况下进行采样，我们也发现该模型在出现漂移之前会先复制语音一段时间，尽管风格相似。它通常在 100-200 个 token 内漂移，并且一旦出现漂移，就会显示出不断扩大的多样性。

为了量化样本中精确记忆出现的频率，我们以 WebText 测试集文章为条件，用 GPT-2 生成样本，并将 GPT-2 生成的重叠率与真实补全的重叠率进行比较。分析结果如下所示，表明 GPT-2 重复训练集中文本的频率低于保留文章的基准频率。

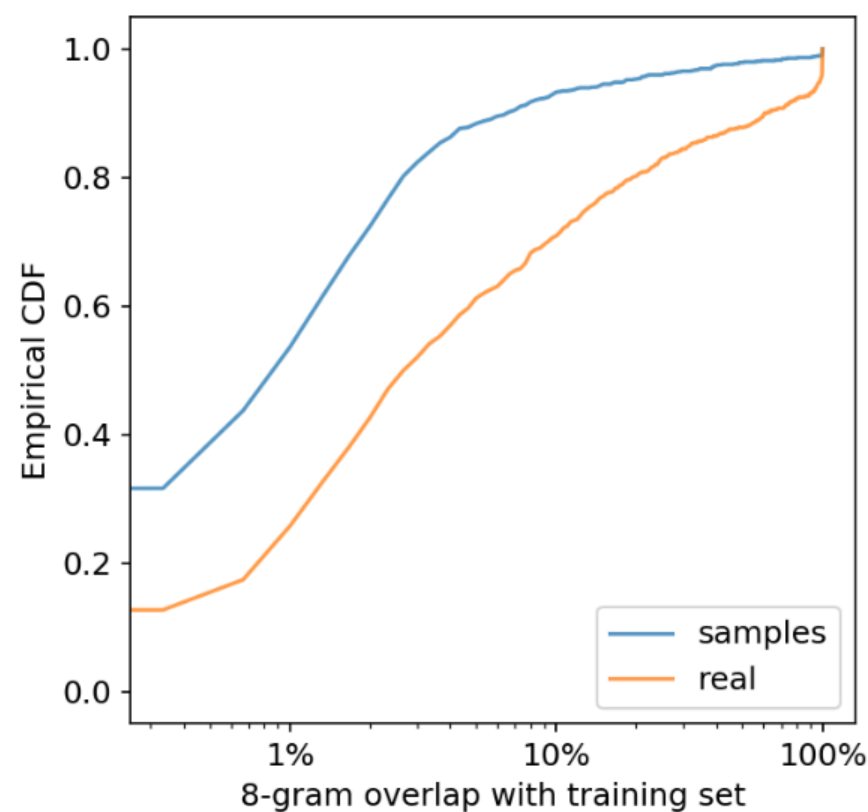


图 5. 8-gram 与 WebText 训练集重叠百分比的累积分布函数 (CDF)，适用于 WebText 测试集和样本（以 WebText 测试集为条件，采用前 k 个截断随机采样， $k = 40$ ）。大多数样本的重叠率低于 1%，其中超过 30% 的样本没有重叠，而测试集的重叠率中值为 2.6%。

8.3. 多样性

表 12 显示了相同随机 WebText 测试集上下文的多次完成，显示了具有标准采样设置的完成的多样性。

8.4. 稳健性

表 13 展示了前面提到的“谈论独角兽”的新闻文章。我们发现该模型能够处理超出分布范围的上下文，但这些样本的质量通常较低。

| 上下文 (WebText 测试) | |
|---|--|
| <p>服装 1：典型</p> <p>这套搭配是我买鞋时想到的第一套。它就像杰克·格兰瑟姆这套造型的夏季版；事实上，我的鞋子和他的耐克 Racer 鞋颜色很接近！我没有穿厚重的哈里斯粗花呢夹克和牛仔衬衫，而是穿了一件棉质 DB 夹克和一件亚麻衬衫。这两种面料（颜色相同）都是夏季的必备单品，因为它们既能搭配深色裤子，也能搭配浅色裤子！正如你所见，它们和深色牛仔裤和鞋子完美搭配。这套造型有点像男装/学院风。</p> <p>总的来说，这身穿搭非常休闲，所以我搭配了一双运动鞋。我不喜欢穿全羊毛西装配运动鞋（就像《GQ》杂志经常展示的那样），但我绝对可以尝试一些休闲的穿搭，比如这双棉质 DB。休闲面料是打造休闲男装的关键，让你的运动鞋更显气质。我甚至建议搭配卡其色斜纹棉布西装或白色亚麻西装。不过，一定要放弃领带，或者穿 T 恤或 Polo 衫；系领带配运动鞋有点太夸张了。</p> | |
| <p>小模型完成</p> <p>.</p> <p>服装 2：高端</p> <p>这套服装简直就是风暴中心。它融合了 ASICS 的街头风格（比如 J. Crew）和 J. Crew 略带传统风格的休闲单品。同样，正如我之前提到的，这是一款低预算单品，而且价格可能非常昂贵。如果你想成为一名休闲运动鞋制造商，这肯定会是一个问题，这没关系，但如果你想尝试他们服装的低价版本，也不要犹豫！</p> <p>这个版本领口很低（这倒也不错），裤腿略长，而且裤腿很长，看起来很漂亮。我喜欢这些衣服的颜色是蓝灰色（这是我的首选颜色，因为它们是浅白色），或者黑色（这是我所有设计中最喜欢的）。和往常一样，我在这家店挑选颜色与款式无关。作为高端产品，我强烈推荐给大家！</p> <p>服装 3：高光</p> <p>这是一个“高光”造型，趁着有机会染点颜色。我觉得颜色是基于</p> | <p>GPT-2 完成</p> <p>为我！</p> <p>服装 2：休闲夹克</p> <p>我知道有些人可能反对穿高领衫配运动鞋，但为了忠于自己，我选择了一件修身的高领衫，颜色跟我当天最喜欢的颜色更接近。总的来说，这是一件经典的高领衫，没有太过花哨。我之前确实把鞋子换成了这双 ASICS Tiger Killshots，真后悔没这么做。鞋子本身很舒服，但我发现配上这双运动鞋会让我的脚看起来更大一些。我会搭配牛仔裤、斜纹棉布裤之类的，但绝对不建议单独穿。因为里面的变数实在太多了。</p> <p>服装 3：连衣裙</p> <p>这是我穿的最后一双 ASICS Tiger Killshots。这双 ASICS 更加现代，也更加流线型。这双鞋的设计是我所能想到的最现代、最简约的！这双鞋用了大量的黑色，看起来很酷。夏天穿它绝对会很舒服。</p> <p>对我来说，搭配西装是拥有一双好运动鞋最简单的方法。差不多就是这样！如果</p> |

表 7. 随机未见上下文（上）以及最小（左）和最大（右）模型的非精选补全。上下文包含 768 个 token，显示约 256 个 token 的段落。补全包含 256 个 token，并已完整显示。生成过程采用 Top-k 随机抽样，k = 40。

| 上下文 (WebText 测试) | |
|--|---|
| 西班牙一些最辉煌的历史景点可以追溯到穆斯林统治时期，包括科尔多瓦清真寺（建于 1897 年），以及同样位于科尔多瓦的阿扎哈拉麦地那（Medina Azahara），后者如今已成废墟，但仍可参观；此外，还有安达卢斯宫殿（建于 1897 年），以及格拉纳达的阿尔罕布拉宫（建于 1897 年），这是一座辉煌完好、保存完好的宫殿。此外，还有两座建于穆斯林西班牙时期的犹太教堂至今仍屹立不倒：托莱多的白色圣玛利亚犹太教堂和老城的科尔多瓦犹太教堂。 | |
| 收复失地运动和帝国时代[编辑] | |
| 小模型完成 | GPT-2 完成 |
| 穆斯林西班牙的时代与世界其他地区截然不同，其中世纪历史相当悠久。西班牙帝国存在于公元前 9 世纪下半叶，但到公元 1150 年，它缓慢地转变为一个小型的世俗化帝国，国家更加繁荣，也更加多元化。公元 1050 年，那不勒斯皇帝自封为那不勒斯皇帝。这至少部分地体现在从该地区西北海岸到西海岸的一系列战争和入侵（例如，多瑙河从直布罗陀一直延伸到西班牙南部，但在同一天被罗马军队迅速摧毁）。在过去的几十年里，他为恢复和维护帝国做出了巨大贡献，其中首先是在西部。他的继任者是费迪南德，他于 1148 年从教皇宝座上夺取了权力，而第一位帝国继承人是瑞典国王阿尔弗雷德二世。帝国再次成为欧洲大陆少数几个独立国家之一。尽管在此期间，其领土上的人口不超过两亿，但民众对皇帝却有很多意见。自 1720 年起，穆斯林占西班牙总人口的一半。 | <p>1092 年至 1095 年的收复失地运动（字面意思是“回归故土”）在伊比利亚半岛建立了卡斯蒂利亚-阿拉贡联合王国，并由穆斯林王朝统治。这是欧洲最后一次将穆斯林统治视为“本土”统治，也是最后一次由基督教徒煽动和组织起来的冲突在欧洲公开露面。</p> <p>穆斯林摩尔人在西班牙的统治，尤其是在非穆斯林群体中，常被称为“格拉纳达征服”，而西班牙有时也被称为“格拉纳达的西班牙”。穆斯林统治的最后几年之后，西班牙迎来了收复失地运动。随着费迪南九世于 1189 年 3 月在普瓦捷战役中的胜利，以及萨拉丁于 1187 年征服耶路撒冷，穆斯林被驱逐或被迫离开伊比利亚半岛。</p> <p>西班牙的收复失地运动时期随着摩尔人北撤而结束。然而，直到收复失地运动时代重新开启，摩尔人的统治才被基督教统治完全取代。收复失地运动从 1229 年持续到 1492 年。</p> <p>西班牙的最后一个穆斯林国家格拉纳达被征服</p> |

表 8. 随机未见上下文（上），以及最小（左）和最大（右）模型的非精选补全。上下文包含 768 个 token，显示约 256 个 token 的段落。补全包含 256 个 token，并已完整显示。生成过程采用 Top-k 随机抽样，k = 40。

| 上下文 (WebText 测试) | |
|---|--|
| <p>Pinky</p> <p>粉色幽灵的 AI 设计与红色幽灵的行为“感觉”相反。粉色幽灵实际上会试图跑到吃豆人前面。这是通过将目标设置在吃豆人当前位置前方 4 个方块处，沿着吃豆人前进的方向来实现的。一个例外是吃豆人向上移动时。由于代码中存在溢出错误，计算中包含的左侧偏移量等于预期的向上偏移量。</p> <p>利用吃豆人的预期位置，玩家可以挖掘 Pinky 的导航代码。这种情况最常见于吃豆人和 Pinky 彼此距离很近的情况。例如，如果它们原本朝着彼此前进，当它们之间的距离小于 4 个方块时，目标方块就位于 Pinky 的后方。如果 Pinky 在这种状态下遇到交叉路口，他会尝试折返（通过转弯）以到达目标方块。</p> <p>玩家经常会在 Pinky 接近十字路口时，通过快速改变方向来做出所谓的“头部假动作”。当 Pinky</p> | |
| <p>小模型完成</p> <p>向左行驶，他们就会得到“正确”的方向。</p> <p>如果该位置实际上与吃豆人之前的位置相距较远，玩家将不会对幽灵的移动做出反应，幽灵将返回到玩家当前的位置。这是因为吃豆人无法通过转弯来移动。</p> <p>幽灵必须前往某个其他地点才能回到起始位置，这可以通过多种方式实现。它通常会直接离开原点，例如，如果吃豆人跳下悬崖，它会回到消失在太空之前的原点。这些幽灵路径的实现方式有很多种，但最明显的一种是，如果玩家不想被困在游戏中，那么最简单的方法就是让吃豆人留在游戏中的时间只有原先的三分之一。</p> | <p>GPT-2 完成</p> <p>即将穿过十字路口时，吃豆人将能够躲避鬼魂的投射物并返回安全屋的安全位置。</p> <p>巴斯特</p> <p>黄色幽灵的 AI 经过特殊设计，试图阻止吃豆人离开安全屋。玩家必须按下空格键打开安全屋的门，然后进入门内，才能找到藏在隔壁房间里的黄色幽灵。</p> <p>玩家还必须注意，如果幽灵看到吃豆人已经在安全屋里，它就不会试图离开。事实上，除非幽灵已经在安全屋里，否则它根本不会尝试离开。</p> <p>这两种行为基于这样的事实：玩家并不总是试图在吃豆人（或 Blinky）之前到达安全屋。</p> <p>如果幽灵（或其他玩家）已经在安全屋内，并在吃豆人或 Blinky 离开后试图离开，玩家也必须注意此故障。此时，幽灵会尝试以类似于“追逐”模式的方式离开安全屋。</p> |
| <p>追逐</p> <p>对于吃豆人来说，这类似于吃豆人在战斗中“追赶”的方式。</p> <p>幽灵可以追逐敌人。这对于吃豆人及其 AI 来说有很多用处。与大多数幽灵游戏不同，这个幽灵只是沿着</p> | <p>吃豆人</p> <p>有些鬼魂会跟着吃豆人。这些鬼魂是</p> |

表 9. 随机未见上下文 (上)，以及最小 (左) 和最大 (右) 模型的非精选补全。上下文包含 768 个 token，显示约 256 个 token 的段落。补全包含 256 个 token，并已完整显示。生成过程采用 Top-k 随机抽样，k = 40。

| | |
|---|---|
| <p>上下文 (WebText 测试)</p> <p>1860 年 11 月，迈克尔·P·戈尔丁下士加入 A 连时，还是一名来自爱尔兰的非技术劳工。戈尔丁在战争中幸存下来。同样来自爱尔兰的帕特里克·奥尼尔下士于 1854 年首次入伍，服役于俄勒冈州美国第三炮兵团 L 连。他返回东海岸，并于 1860 年加入该连队。奥尼尔服役至 1874 年，之后被任命为纽约州威利茨角国家公墓的主管。本杰明·布朗下士是来自纽约州奥兰治县的一名鞋匠。1862 年 8 月，他加入了新成立的第 124 纽约志愿军团，并于同年 10 月成为 61 名转入 A 连的士兵之一。1864 年 2 月，布朗再次加入该连队，当时该连队驻扎在布兰迪车站。1867 年服完兵役后，他回归平民生活。</p> <p>6 月 10 日，技工威廉·柯林斯晋升为下士，可能是为了填补渡詹姆斯河作战指挥的空缺。柯林斯的服役记录并未反映他获得晋升所需的素质，但他显然克服了一些严重的问题。柯林斯出生于纽约州萨克茨港，1853 年 12 月，22 岁的他入伍，并于 1858 年 12 月再次入伍。就在 1861 年 4 月战争爆发前一个月，柯林斯“堕落”了，直到三年后才被抓获。1864 年 3 月 22 日，他回到了连队，并接受了审判</p> | |
| <p>小模型完成</p> <p>因军事指控被美国军队开除。他回到公司，并在内战期间被分配到内务部。</p> <p>作为一名工程师，柯林斯也在他的家乡纽约接受过工程学培训。在联合工程学院完成研究生学业后，他在美国陆军服役了两年。在那里，他为丹尼尔·菲弗上尉效力，并于 1870 年 5 月晋升为上校，菲弗上尉入伍三年后。在第一师完成第一年后，柯林斯回到英国，在陆军工程兵团服役两年，到第二年，他开始体验服役的乐趣。—endoftext— 莫斯科，7 月 5 日（路透社）——英国的俄罗斯官方媒体、政府消息人士和专家表示，莫斯科试图淡化叙利亚反对派的“军事集结”，同时继续支持唐纳德·特朗普总统在美墨边境修建隔离墙的计划，称其为“军事政变”。</p> <p>美国领导的叙利亚战争老兵和叙利亚特使迈克尔·弗林（右），右，与美国叙利亚政治分析家艾哈迈德·科里亚特打手势，</p> | <p>GPT-2 完成</p> <p>同年 6 月，他因逃兵罪被判处枪决，但仍继续服役。</p> <p>詹姆斯·J·达尔上尉于 1829 年独立战争期间出生于弗吉尼亚，1836 年随父亲参军。他曾在弗吉尼亚第二步兵团服役至 1845 年，之后转入西部预备役团；退伍后，他重新加入了纽约第二十五志愿军团第一连（1845 年该团约有 4,500 人）。他的兄弟詹姆斯也在同一旅的同一连队中担任同一职务。该团一直驻扎在康涅狄格州直到 1846 年，达尔上尉退伍前曾在第一连服役不同时期。他参加过墨西哥战争，曾在圣安东尼奥担任约翰·沙利文准将的参谋两年。1863 年半岛战役期间，他指挥一个 23 人的连队，负责指挥拥有 54 门重型迫击炮的炮台和炮台的炮艇“平塔”号、一艘炮艇和其他几艘船只。1865 年秋，达尔上尉在第一连服役。两年后，他晋升为少尉。</p> <p>约翰·A·奥船长</p> |

表 10. 随机未见上下文（上），以及最小（左）和最大（右）模型的非精选补全。上下文包含 768 个 token，显示约 256 个 token 的段落。补全包含 256 个 token，并已完整显示。生成过程采用 Top-k 随机抽样，k = 40。

| 上下文 (WebText 测试) | |
|--|--|
| <p>Xmonad、Xresources 和字体</p> <p>Xresources 确实很麻烦（对我来说就是这样）。所以，以下是一些帮助：</p> <p>首先，Xmonad 通常通过 XFT 来处理字体。Xft 是 X FreeType 接口库，用于处理 Xmonad/urxvt 中字体的实际渲染（Xresources 为在 X 窗口管理器下启动的程序指定了一组资源。各个程序的设置在 Xresources 中指定，如下所示：Urxvt.background:background settins here）。</p> <p>无论如何，在 Xresources 中使用此语法设置字体：</p> <p>urxvt*font:xft:你的字体名称放在这里:size=字体的大小放在这里,xft:fall-back 字体 1 放在这里</p> <p>所以，这是通过 Xresources 指定的 urxvt 中字体的最基本部分。</p> <p>修复 Cygwin 中的 /.ssh/ 权限</p> <p>在 Windows 上的 Cygwin 中，我发现无法将 /.ssh/ 文件夹的权限设置为 0600，而 ssh 需要该权限才能允许你使用密钥。我的症状是，无论我做什么，它总是会修改所有者和组。</p> | |
| <p>小模型完成</p> <p>之前设置成了 /.ssh/，情况就是这样。这个问题无法完全解决。我发现，只需用鼠标输入用户名，然后尝试将组的权限设置为“0600”，这样就更容易解决了，我的系统现在就是这么配置的。</p> <p>然而，自从 Linux 上的 Cygwin 7 开始，我发现这样做对我的系统来说已经“没有必要”了！我还安装了一个名为 chroot 的包。它允许你安装/升级到你的系统目录。</p> <p>因此，从 Windows 8 及更高版本开始，安装该软件包可能需要 chroot，所以我反复尝试，发现在 Linux 上安装并卸载软件包时问题并不大。截至我上次撰写 Cygwin 文章时，该脚本已经完成了它的工作（sudo 命令完成了剩下的工作），如果您尝试这样做，则需要记录更改。如果这还不够让人困惑的话，我尝试了一下，然后回到我的 /.ssh/ 目录和 chroot 脚本。我发现权限终于用完了（因为我忘了从文件系统创建 ssh 的符号链接，而且</p> | <p>GPT-2 完成</p> <p>，使这些文件变为只读。我尝试过直接启动 ssh，并使用类似 passwd 的程序将文件所有者/组权限修改为 0644，但没用。这很烦人，但也没那么糟糕，只是不得不费劲地摆弄权限而已。</p> <p>所以我决定反过来做：</p> <p>对于 Cygwin 目录中的每个 .ssh 文件夹，将每个文件和目录的权限设置为 2644。然后使用以下命令启动 ssh： ssh-add /.ssh/ 第一次连接时，您在 Cygwin 中刚刚拥有的 /.ssh 文件夹的权限将设置为 0600。然后，每次再次连接时，您的权限将更改为 0644。</p> <p>/.xsessionrc 中使用的 Xresources 的 Makefile</p> <p>这些是在我的 .xsessionrc 中制作的：</p> <p>.xprofile: .xsession 错误：颜色=" "</p> <p>例如：</p> <p>.xprofile: 颜色=蓝色</p> <p>由于颜色存储在单独的文件中，因此 /.xsession-errors 中的颜色看起来是青色而不是蓝色。 这主要是因为，截至目前，只有颜色分配给你的</p> |

表 11. 随机未见上下文（上图）以及最小（左图）和最大（右图）模型的非精选补全。上下文包含 768 个 token，显示约 256 个 token 的段落。补全包含 256 个 token，并已完整显示。生成过程采用 Top-k 随机抽样，k = 40。

| 上下文 (WebText 测试) | | |
|---|---|--|
| <p>Aaron 喜欢薄荷巧克力蛋糕，但他要求搭配迷你巧克力碎，所以我在蛋糕夹层之间撒了一些。我手边还有一些 Peppermint Jo Jos，所以我把它们碾碎，和一些碎蛋白饼一起撒了进去，何乐而不为呢？这简直就是薄荷巧克力碎饼干香脆可口的盛宴。我没有称量每种配料的用量，但尝过成品后，我真希望自己加得更多。你可以添加任何你想加的东西——碎拐杖糖、薄荷皮等等。别害怕用力过猛。口感 = 不错。</p> | | |
| <p>如果你没有 7 英寸的蛋糕盘，你可以用这个蛋糕盘制作 3 个较短的 8 英寸蛋糕层</p> | | |
| <p>GPT-2 完成 1</p> <p>食谱。我做了长一点的版本，然后用烤盘做了一层奶油糖霜，并将三分之一放进冰箱作为糖霜装饰。一定要在做这个蛋糕之前准备好所有糖霜，这样你就不用等蛋白都粘在搅拌碗底部了。</p> <p>你想把蛋白打发到硬性发泡吗？那就把蛋糕盘放在一锅沸水上，加热至微温。这不需要超过 3-5 分钟；刚好够开始搅拌。</p> | <p>GPT-2 完成 2</p> <p>食谱。如果你知道，并且知道如何制作那个蛋糕，那么你的生活将会焕然一新。</p> <p>说到饼干屑或巧克力片，我有两条非常简单的规则。1) 我买的原糖，要等它冷却下来，上面至少沾上半茶匙的糖才会用。这是我的规则。2) 如果我的食谱需要某种材料，而我想用它做菜，我不会用网上买的其他材料代替。就像这样。</p> <p>我发现这张卡片非常可爱，上面有可爱的小猫</p> | <p>GPT-2 完成 3</p> <p>配方，可能会产生大约相同数量的蛋糕。</p> <p>巧克力酪乳蛋糕 打印 巧克力蛋糕块，裹上酪乳，撒上酥脆的巧克力碎，淋上巧克力酱。简单却不失为巧克力蛋糕的绝妙之选！作者：妈妈做的 巧克力酪乳蛋糕的配料：1 1/2 杯无盐黄油，室温</p> <p>1 1/2 杯砂糖 3 个大鸡蛋加 1 个蛋黄 巧克力釉料 1/3 杯可可粉 1 3/4 杯糖粉 6 盎司半</p> |
| <p>GPT-2 完成 4</p> <p>食谱。</p> <p>蛋白酥饼</p> <p>1 3/4 杯黄油，软化 1 杯糖 1 个蛋黄 3 汤匙浓奶油 3 1/2-4 杯面粉 少许盐 薄荷乔乔酱 1 1/4 杯糖粉 1 杯切碎的山核桃 1/2 杯切碎的薄荷叶 1/2 杯切碎的新鲜薄荷 (约 1/2 片) 1 茶匙薄荷提取物 1/4 茶匙海盐 少许海盐</p> | <p>GPT-2 完成 5</p> <p>食谱。</p> <p>圣诞节早上我的两次失败并非因为糖霜技术不好。那是另一个成功的故事。蛋糕湿润得恰到好处，嗯，就是一个漂亮的蛋糕。蛋糕里几乎没有空气和水分，但巧克力和薄荷完美地融合在一起。</p> <p>有一件事让我有点尴尬，那就是所有这些材料——蛋黄、糖、融化的黄油、巧克力碎、融化的巧克力和薄荷——都必须直接拌进蛋糕糊里，这非常耗时。一位朋友建议，如果我把蛋糕糊放在一个单独的碗里，我</p> | <p>GPT-2 完成 6</p> <p>这个食谱可以做出 6-9 份美味的巧克力蛋糕。为了防止蛋糕层粘连，我用了一个 8 英寸的圆形挞模，铺上烘焙纸，然后切出 4 个 7 英寸的圆形蛋糕，把蛋糕层夹在中间（这个方法也适用于单层巧克力蛋糕）。准备烘烤蛋糕时，取下烘焙纸，将烤箱预热至 350 华氏度 (约 175 摄氏度)，烘烤约 30-40 分钟，直到蛋糕表面呈金黄色，取出时干净为止。</p> <p>巧克力酪乳夹心蛋糕配薄荷乔</p> |

表 12. GPT-2 基于相同上下文生成的非 Cherry-Pick 补全 (来自 WebText 测试)。上下文包含 384 个词条 (显示为截断)，生成词条数量为 128 个。生成过程中采用 Top-k 随机抽样，k = 40。

| |
|--|
| <p>语境（人工书写）：科学家发现了一个令人震惊的现象：一群独角兽生活在安第斯山脉一个偏远、从未有人探索过的山谷里。更让研究人员惊讶的是，这些独角兽竟然能说一口流利的英语。</p> <p>GPT-2：科学家根据这些独角的独特形状，将其命名为“奥维德独角兽”。这些长着四角、银白色的独角兽此前在科学界尚不为人所知。</p> |
| <p>如今，经过近两个世纪，引发这一奇怪现象的谜团终于被解开了。</p> <p>拉巴斯大学进化生物学家豪尔赫·佩雷斯博士和几位同伴在安第斯山脉探险时，发现了一个小山谷，那里没有其他动物或人类。佩雷斯注意到，山谷里似乎有一个天然喷泉，周围环绕着两座岩石山峰和银色的积雪。</p> |
| <p>佩雷兹和其他人随后继续深入山谷。“当我们到达一座山峰的顶峰时，水看起来是蓝色的，上面还有一些水晶，”佩雷兹说。</p> <p>佩雷斯和他的朋友们看到这群独角兽，都惊呆了。这些生物离我们如此之近，以至于触手可及。从空中就能看到它们，不用走太远就能看到。</p> |
| <p>在研究这些奇异生物时，科学家们发现它们还会说一些相当标准的英语。佩雷兹说道：“例如，我们可以看到它们有一种共同的‘语言’，类似于方言或方言学。”</p> |
| <p>佩雷斯博士认为，独角兽可能起源于阿根廷，那里的人们认为这些动物是人类到达南美洲之前生活在那里的那个失落种族的后代。</p> |
| <p>虽然它们的起源尚不清楚，但有些人认为，这些生物或许是在人类文明出现之前的某个时期，人类与独角兽相遇时产生的。佩雷兹说：“在南美洲，这种事似乎很常见。”</p> |
| <p>然而，佩雷兹也指出，唯一能确定独角兽是否是失落外星种族后裔的方法很可能是通过 DNA。“但它们似乎能够很好地用英语交流，我认为这是进化的标志，或者至少是社会组织发生变化的标志，”这位科学家说。</p> |

表 13. GPT-2 在分布外上下文上的条件生成。从 k = 40 生成的 10 个样本中挑选出来。

| | |
|--|---|
| <p>文章：36000 年前，史前人类在如今法国的一个洞穴粗糙的石灰岩壁上描绘了一系列令人难以置信的史前野兽。如今，借助尖端技术，肖维岩洞中的这些艺术作品得以重现，打造出世界上最大的复制洞穴。这座名为“蓬达尔克洞穴”（Caverne du Pont-d'Arc）的人工洞穴距离原址几英里，位于法国南部的瓦隆蓬达尔克，内有 1000 幅精心复制的图画，以及约 450 块骨骼和其他遗迹.....</p> <p>36,000 年前，穴居人和妇女在洞穴粗糙的石灰岩壁上描绘了一系列令人难以置信的史前野兽，现在他们制作了一个复制品（如图）</p> <p>...</p> | |
| <p>GPT-2：位于法国南部瓦隆蓬达尔克的原址已被联合国教科文组织列为世界遗产，也是已知最古老、保存最完好的人工洞穴。复制洞穴距离原址几英里，位于法国南部瓦隆蓬达尔克。洞穴内有 14 种不同动物的图像，包括披毛犀、猛犸象和大型猫科动物。</p> | <p>参考：该洞穴模仿了法国著名的蓬达尔克洞穴（Caverne du Pont-d'Arc），这是人类建造的最古老、保存最完好的洞穴。复制品包含全部 1000 幅壁画，其中包括 425 幅壁画，例如披毛犀和猛犸象。微小的细节通过 3D 建模和变形技术复制，这些技术通常用于拍摄宽屏图像。这座现代洞穴还保留了原始洞穴中保存的熊爪印、骨骼和细节的复制品。</p> |
| <p>文章：俄罗斯石油亿万富翁罗曼·阿布拉莫维奇曾说：“如果你能找到我的潜艇，它就是你的。”这位寡头兼切尔西足球俱乐部老板的潜艇只是他价值 3 亿英镑的超级游艇“日蚀号”（Eclipse）的额外配置之一（非常适合不被发现地登陆），看来他并非唯一一个在水下探险上投入巨资的富商。</p> <p>海洋被称为地球真正的“最后边疆”，但其探索程度却依然有限，以至于亿万富翁们纷纷排队购买船只，以便一睹世界海洋中壮丽的海景和令人惊叹的野生动物。所以，如果你银行账户里有几百万美元的闲钱，想要拍一些度假照片和 Instagram 帖子，让你在 Instagram 上爆红，不妨入手一艘这样的潜水艇.....畅游鲸鱼：OrcaSub 潜水艇将带你潜入 2000 英尺深的深海，在两个加压的有机玻璃观景穹顶中进行最佳探索。</p> <p>...</p> | |
| <p>GPT-2：U-Boat Worx 潜艇可下潜至 984 英尺（300 米），并能够潜入海底。它可以用于潜水、浮潜，甚至浮潜。该潜艇还配备了浮潜平台。</p> | <p>参考：体验一下拥有最先进潜水艇的詹姆斯·邦德反派。有些潜水艇可以从游艇改装成潜艇，有些则可以脱离潜艇进行快速探索。奥利弗旅行社 (Oliver's Travels) 提供“Mile Low”潜水艇套餐，包含厨师和管家服务，每晚价格为 17.5 万英镑。</p> |
| <p>文章：阿米娜·阿里·卡西姆坐在她腿上，最小的孙女正坐在她腿上，用头巾擦着眼泪。这个只有几个月大的女婴，正是她在空袭开始的那天晚上拼命捂住耳朵的。她说，在也门比里姆岛的一座乡村清真寺里，她躺在床上辗转反侧，数着婴儿啼哭时传来的爆炸声。</p> <p>但情况本来可能更糟。第一枚导弹落下时，他们可能还在屋里。“邻居冲我丈夫喊：‘你得走，他们来了。’我们就跑。我们刚离开房子，第一枚导弹就落在了它旁边，然后第二枚也落在了它身上。它把一切都烧成了灰烬，”卡西姆告诉我们。</p> <p>...</p> | |
| <p>GPT-2：也门正处于内战之中。沙特阿拉伯正在领导联军的轰炸行动。他们已经对也门进行了两个多月的轰炸。</p> | <p>参考：阿米娜·阿里·卡西姆一家在逃离也门前，曾在一座清真寺寻求庇护。成千上万像他们一样的人正乘船前往吉布提。沙特阿拉伯一直在对也门进行猛烈打击，试图击败胡塞叛军。</p> |

表 14. GPT-2 生成的摘要以及 CNN 和每日邮报数据集测试集上的参考摘要。

| | |
|---|--|
| <p>英文参考</p> <p>这张重新发行的专辑名为《The Next Day Extra》，以三张光盘的形式呈现：原始专辑、未发布的录音室录音和混音，以及一张包含已经发布的四个剪辑的 DVD。</p> | <p>GPT-2 法语翻译</p> <p>Les nouvelles 重新发行，推出首张唱片、专辑、录音室录音和混音版，以及 el"ecran de quelques 剪辑的 DVD。</p> |
| <p>法国参考</p> <p>一个明确的疝气功能的人可以利用前卫的方法来识别疝气的功能，我的疝气功能是我的，我的。</p> | <p>GPT-2 英语翻译</p> <p>该版本最初以三张光盘的形式发行：原版专辑、录音室版的编辑和混音版、以及包含前一天发行的四个剪辑的 DVD。</p> |
| <p>英文参考</p> <p>一名男子解释说，他接受的免费疝气手术将使他能够再次工作。</p> | <p>GPT-2 法语翻译</p> <p>一个明确的疝气功能的人可以利用前卫的方法来识别疝气的功能，我的疝气功能是我的，我的。</p> |
| <p>法国参考</p> <p>Un homme a expliqué que l'action gratuite qu'il avait subie pour soigner une hernie lui permettrait de travailler.</p> | <p>GPT-2 英语翻译</p> <p>一名男子告诉我，他被承诺的手术酬金不足以让他去旅行。</p> |
| <p>英文参考</p> <p>克里在伦敦通过视频连线举行的一次政府公开会议上表示：“毫无疑问，总统和我以及政府其他官员实际上已经了解到一些自动驾驶仪上发生的事情，因为这种能力一直存在，可以追溯到第二次世界大战和冷战时期的艰难岁月，当然还有 9/11</p> | <p>GPT-2 法语翻译</p> <p>” Le Premier Conference de l' édition Internationale en ligne sur la vide, le ministre de l' 'Edition Internationale a ´et ´e vide, lui dit : “Il n'y a qu'une question que le le 'esident et moi et tous les autres personnes d'une Administration qui a ´et ´e ´evident par la fois, par la part 战争部分</p> |
| <p>法国参考</p> <p>在政府间会议上发言 克里先生在伦敦通过视频连线发表讲话时说：“毫无疑问，总统、我和政府其他成员在自动驾驶过程中学到了某些东西，因为我们有机会，从第二次世界大战开始，一直到冷战的艰难岁月，当然还有 9·11 事件。”</p> | <p>GPT-2 英语翻译</p> <p>在美国和伦敦举行的一次会议上，国务卿约翰·克里表示：“不可否认的是，总统、我本人以及政府其他官员都知道，为了能够以更加自动化的方式完成某些事情，我们过去做出了某些选择。”</p> |

表 15. GPT-2 生成的英语到法语和法语到英语的翻译。

上下文（文章和之前的问题/答案对）

2008 年夏季奥运会火炬传递活动于 2008 年 3 月 24 日至 8 月 8 日在 2008 年夏季奥运会开幕前举行，主题为“同一个世界，同一个梦想”。传递计划于 2007 年 4 月 26 日在中国北京宣布。此次传递活动也被组织者称为“和谐之旅”，历时 129 天，火炬传递距离达 137,000 公里（85,000 英里），是自 1936 年夏季奥运会开始火炬传递以来，传递距离最长的一次。

火炬于 3 月 24 日在奥运会发源地希腊奥林匹亚点燃后，经雅典帕纳辛奈克体育场，再于 3 月 31 日抵达北京。从北京出发，火炬传递的路线穿越六大洲。火炬曾访问过丝绸之路沿线城市，象征着中国与世界各国之间古老的联系。此次传递还包括从中国境内攀登位于尼泊尔和中国西藏边境的珠穆朗玛峰，珠穆朗玛峰的中方通道因此次活动而特别关闭。

问：主题是什么？ 答：“同一个世界，同一个梦想”。

问：比赛全程多长？ 答：137,000 公里

问：它比以前的更大吗？ 答：没有。

问：比赛在哪里开始？ 答：希腊奥林匹亚

问：那个地方有什么值得一提的吗？ 答：奥运会的发源地

问：他们后来去了哪里？ 答：雅典

问：比赛持续了几天？ 答：七天

问：他们参观了什么著名地标吗？ 答：帕纳辛奈克体育场

问：他们爬过什么山吗？ 答：

典型答案：珠穆朗玛峰
Turker 回答：未知，是的，是的，是的

表 16. 选定的 CoQA 完成。

上下文（文章和之前的问题/答案对）

汤姆总是和 54 岁的秘书凯瑟琳·格林形影不离。他会在她办公室里走来走去，还会和她一起去购物。“大多数人似乎并不介意汤姆，”凯瑟琳说，她觉得汤姆很棒。“他是我的第四个孩子，”她说。她或许会把他当成自己的儿子一样看待。他四处奔波，买食物，支付医疗费用和税款，但实际上汤姆是一只狗。
凯瑟琳和汤姆住在瑞典。在这个国家，每个人都应该按照政府制定的规则过着有序的生活，政府也为民众提供高水平的护理。这种水平的护理需要花钱。

瑞典人什么都要纳税，所以养狗意味着要交更多税也就不足为奇了。有些人为了养狗的权利，每年要缴纳高达 500 瑞典克朗的税款，这笔钱由政府用于狗狗医院的治疗，有时也用于狗狗生病时的医疗。然而，大多数此类治疗费用昂贵，因此主人往往选择为狗狗的健康甚至生命付出代价。

在瑞典，狗主人必须为其爱犬造成的任何损害赔偿。瑞典养犬俱乐部的一位官员解释了这意味着什么：如果你的爱犬在路上跑，被过往车辆撞到，作为主人，你必须赔偿车辆造成的任何损害，即使你的爱犬在事故中丧生。

问：凯瑟琳多大了？答：
54 岁

问：她住在哪里？答：

典型答案：斯德哥尔摩
特克回答：瑞典，瑞典，在瑞典，瑞典

表 17. 选定的 CoQA 完成。