GPT-4技术报告

OpenAI*

摘要

我们报告了GPT–4的开发，这是一个大规模的多模态模型，可以接受图像和文本输入并产生文本输出。虽然在许多现实世界的场景中不如人类，但GPT–4在各种专业和学术基准上表现出人类水平的表现，包括以大约前10%的分数通过模拟律师考试。GPT–4是一个基于Transformer model的模型，经过预训练，可以预测文档中的下一个令牌。培训后的调整过程提高了真实性和对期望行为的遵守程度。这个项目的一个核心组成部分是开发基础设施和优化方法，这些方法在广泛的规模范围内表现得可预测。这使我们能够根据不超过GPT 4计算量1/1000的模型准确预测GPT 4性能的某些方面。

1 导言

本技术报告介绍了GPT–4，这是一个大型多模态模型，能够处理图像和文本输入并产生文本输出。这种模型是一个重要的研究领域，因为它们具有广泛应用的潜力，如对话系统、文本摘要和机器翻译。因此，近年来，它们一直是人们极大兴趣和进步的主题【1–28】。

开发这种模型的主要目标之一是提高它们理解和生成自然语言文本的能力，特别是在更复杂和微妙的场景中。为了测试它在这种情况下的能力，GPT–4在最初为人类设计的各种考试中进行了评估。在这些评估中，它表现得相当好，通常得分超过绝大多数人类考生。例如，在一次模拟律师考试中，GPT–4的分数在考生中排名前10%。这与GPT的3.5分形成鲜明对比，后者排名倒数10%。

在一套传统的NLP基准测试中，GPT–4优于以前的大型语言模型和大多数最先进的系统（通常有特定于基准测试的训练或手工工程）。在MMLU基准【29，30】上，一套涵盖57个科目的英语多项选择题，GPT–4不仅在英语方面远远超过现有模型，而且在其他语言方面也表现强劲。在MMLU的翻译版本上，GPT–4在26种语言中的24种超过了英语的最先进水平。我们将在后面的章节中更详细地讨论这些模型功能结果，以及模型安全性改进和结果。

该报告还讨论了该项目的一个关键挑战，即开发深度学习基础设施和优化方法，这些方法在广泛的规模上表现可预测。这使我们能够预测GPT–4的预期性能（基于以类似方式训练的小跑步），并在最后一次跑步中进行测试，以增加我们训练的信心。

尽管GPT–4有其功能，但它与早期的GPT模型有类似的局限性[1，31，32]：它不完全可靠（例如，可能出现“幻觉”），具有有限的上下文窗口，并且不学习

*请将此作品引用为“OpenAI（2023）”。完整的作者贡献声明出现在文件的末尾。

凭经验。使用GPT–4的输出时应小心，尤其是在可靠性很重要的情况下。

GPT–4的能力和局限性带来了重大和新颖的安全挑战，鉴于潜在的社会影响，我们相信仔细研究这些挑战是一个重要的研究领域。该报告包括一个广泛的系统卡（在附录之后），描述了我们预见的偏见、虚假信息、过度依赖、隐私、网络安全、扩散等方面的一些风险。它还描述了我们为减轻GPT–4部署的潜在危害而采取的干预措施，包括与领域专家的对抗性测试，以及模型辅助的安全管道。

2 本技术报告的范围和限制

本报告重点介绍GPT–4的能力、局限性和安全特性。GPT–4是一种Transformer model风格的模型【33】，使用公开可用的数据（如互联网数据）和第三方提供商许可的数据，预先训练以预测文档中的下一个令牌。然后使用来自人类反馈的强化学习（RLHF）对该模型进行微调【34】。鉴于竞争格局和GPT–4等大规模模型的安全影响，本报告不包含有关架构（包括模型大小）、硬件、训练计算、数据集构建、训练方法或类似内容的更多细节。

我们致力于对我们的技术进行独立审计，并在本版本随附的系统卡中分享了该领域的一些初步步骤和想法。2我们计划向更多第三方提供进一步的技术细节，这些第三方可以建议我们如何权衡上述竞争和安全因素与进一步透明的科学价值。

3 可预测的缩放

GPT–4项目的一大重点是建立一个可预测扩展的深度学习堆栈。主要原因是，对于像GPT–4这样的非常大的训练运行，进行广泛的特定于模型的调整是不可行的。为了解决这个问题，我们开发了基础设施和优化方法，这些方法在多个规模上具有非常可预测的行为。这些改进使我们能够可靠地预测GPT–4性能的某些方面，这些性能来自使用1,000 ×–10,000 ×较少计算训练的较小模型。

3.1 损耗预测

经过适当训练的大型语言模型的最终损失被认为很好地近似于用于训练模型的计算量的幂律【35，36，2，14，15】。

为了验证我们的优化基础设施的可扩展性，我们通过拟合具有不可约损失项的标度律（如Henighan等人【15】）来预测GPT–4在我们内部代码库（不是训练集的一部分）上的最终损失：L（C）=aCb+C，来自使用相同方法训练的模型，但使用的计算量最多比GPT–4少10，000倍。这一预测是在运行开始后不久做出的，没有使用任何部分结果。拟合的标度律高精度地预测了GPT 4号的最终损耗（图1）。

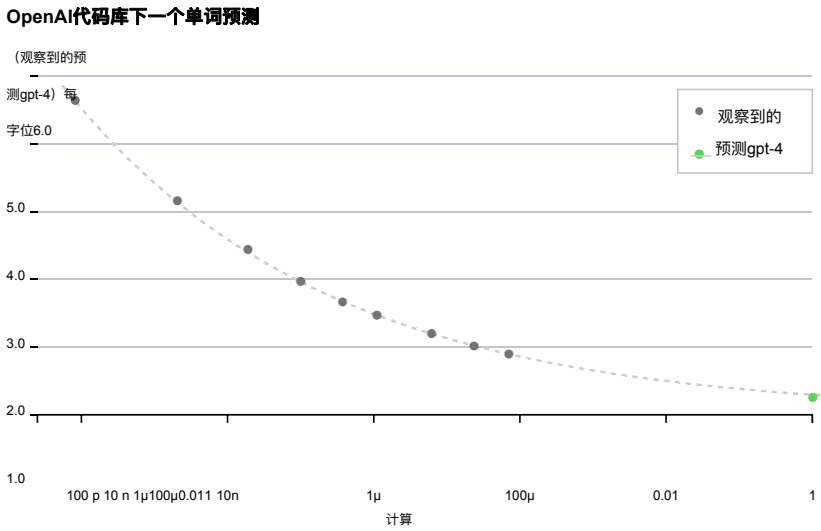
3.2 HumanEval上能力的扩展

在训练前对模型的能力有所了解可以改进围绕一致性、安全性和部署的决策。除了预测最终损失，我们还开发了一种方法来预测更可解释的能力指标。一个这样的指标是HumanEval数据集【37】的通过率，它衡量综合不同复杂性的Python函数的能力。我们成功地预测了HumanEval数据集子集的通过率，方法是从最多减少1000倍计算的模型中进行外推（图2）。

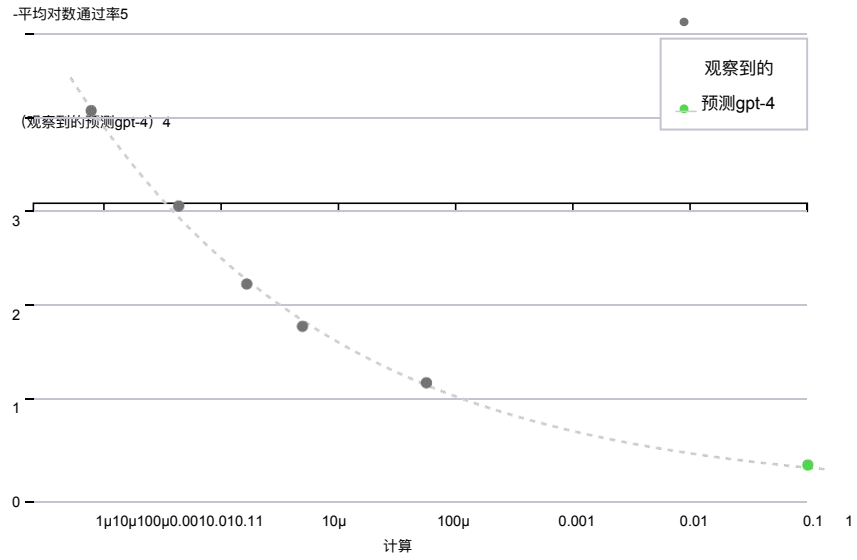
对于HumanEval中的单个问题，性能偶尔会随着规模的扩大而恶化。尽管存在这些挑战，我们还是

找到了一个近似的幂律关系–EP[log(pass_rate(C))]=αΣC–k

2除了随附的系统卡片，OpenAI将很快发布关于人工智能系统的社会和经济影响的其他想法，包括有效监管的必要性。



图一。GPT-4和更小型号的性能。指标是从我们的内部代码库派生的数据集的最终损失。这是一个方便的大型代码令牌数据集，不包含在训练集中。我们选择关注损失，因为在不同数量的训练计算中，它往往比其他测量方法噪声更小。虚线显示了适合较小模型（不包括GPT-4）的幂律；这种拟合准确地预测了GPT 4号的最终损失。x轴被训练计算归一化，使得GPT-4为1。



图二。GPT-4和更小型号的性能。指标是HumanEval数据集子集的平均对数通过率。虚线显示了适合较小模型（不包括GPT-4）的幂律；这种拟合准确地预测了GPT-4的性能。x轴被训练计算归一化，使得GPT-4为1。

其中k和α是正常数，P是数据集中问题的子集。我们假设这种关系适用于该数据集中的所有问题。在实践中，很低的通过率很难或不可能估计，所以我们限制问题P和模型M，使得给定一些大的样本预算，每个问题由每个模型至少解决一次。

我们在训练结束前，仅使用训练前可用的信息，在HumanEval上记录了对GPT–4表现的预测。根据较小模型的表现，除了15个最难的人类评估问题之外，所有问题都被分成6个难度桶。第三个最简单的桶的结果如图2所示，表明对于HumanEval问题的这个子集，结果预测非常准确，我们可以准确地估计几个较小模型的log（pass_rate）。对其他五个桶的预测表现几乎一样好，主要的例外是GPT–4不如我们对最容易的桶的预测。

某些能力仍然难以预测。例如，逆标度奖【38】提出了几个模型性能随标度而降低的任务。与魏等人最近的研究结果相似。[39]，我们发现GPT–4逆转了这一趋势，如图3中一项叫做后见之明忽视的任务[40]所示。

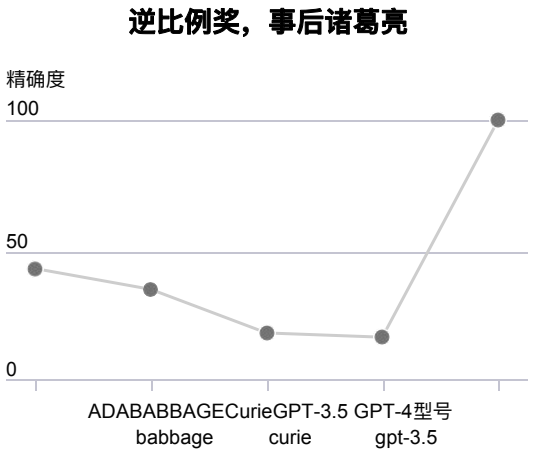


图3。GPT-4和更小型号在后见之明忽略任务中的表现。精度显示在y轴上，越高越好。ada、babbage和curie指的是通过OpenAI API【41】提供的模型。

我们认为，准确预测未来能力对安全非常重要。展望未来，我们计划在大型模型训练开始之前改进这些方法，并跨各种功能注册性能预测，我们希望这成为该领域的共同目标。

4 能力

我们在一系列不同的基准上测试了GPT–4，包括最初为人类设计的模拟考试。3我们没有为这些考试做专门的培训。考试中的少数问题是模型在训练中看到的；对于每次考试，我们运行一个删除这些问题的变体，并报告两个问题中较低的分数。我们认为结果具有代表性。有关污染的更多详细信息（方法和每次检查的统计数据），请参见附录C。

考试来源于公开的材料。考试问题包括选择题和自由回答题；我们为每种格式设计了单独的提示，并且图像包含在需要它的问题的输入中。评估设置是根据一组验证考试的表现设计的，我们报告延期考试的最终结果。总分数是通过使用公开可用的方法结合每次考试的多项选择和自由回答问题分数来确定的。有关考试评估方法的更多详细信息，请参见附录A。

3我们使用训练后的RLHF模型进行这些检查。

考试	GPT-4	GPT-4（无视力）	GPT-3.5
统一律师考试（MBE+MEE+MPT）	298/400（约90）	298/400（约90）	213/400（第10位）
LSAT	163（第88位）	161（第83位）	第149次（第40次）
SAT循证读写	710/800（约93路）	710/800（约93路）	670/800（第87位）
SAT数学	700/800（第89位）	690/800（第89位）	590/800（第70次）
研究生入学考试（GRE）定量	163/170（第80次）	157/170(~62)	147/170（第25次）
研究生入学考试（GRE）口语	169/170（第99次）	165/170（第96次）	154/170（约63）
研究生入学考试（GRE）写作	4/6（第54位）	4/6（第54位）	4/6（第54位）
USABO半决赛2020	87/150（第99-100次）	87/150（第99-100次）	43/150（31-33）
2022年USNCO地方科考试	36/60	38/60	24/60
医学知识自我评估计划	75%	75%	53%
Codeforces评级	392（低于第5名）	392（低于第5名）	260（低于第5名）
AP艺术史	5（第86-100次）	5（第86-100次）	5（第86-100次）
AP生物学	5（第85-100次）	5（第85-100次）	4（第62-85次）
微积分	第4（第43-59）	第4（第43-59）	1（第0-7次）
AP化学	4（第71-88）	4（第71-88）	2（第22-46次）
AP英语语言与写作	2（第14-44次）	2（第14-44次）	2（第14-44次）
AP英语文学与写作	2（8-22）	2（8-22）	2（8-22）
AP环境科学	5（第91-100）	5（第91-100）	5（第91-100）
AP宏观经济学	5（第84-100次）	5（第84-100次）	第2（第33-48）
微观经济学	第5（第82-100次）	4（第60-82）	4（第60-82）
AP物理2	4（第66-84次）	4（第66-84次）	3（第30-66次）
AP心理学	第5次（第83-100次）	第5次（第83-100次）	第5次（第83-100次）
AP统计	5（第85-100次）	5（第85-100次）	3（第40-63）
美联社美国政府	5（第88-100次）	5（第88-100次）	4（第77-88次）
美联社美国历史	5（第89-100次）	4（第74-89次）	4（第74-89次）
AP世界历史	4（第65-87次）	4（第65-87次）	4（第65-87次）
AMC 10	30/150（第6-12次）	36/150（第10-19次）	36/150（第10-19次）
AMC 12	60/150（第45-66次）	48/150（第19-40次）	30/150（第4-8次）
品酒师入门（理论知识）	92%	92%	80%
注册侍酒师（理论知识）	86%	86%	58%
高级侍酒师（理论知识）	77%	77%	46%
Leetcode（简易）	31/41	31/41	12/41

Leetcode（中等）	21/80	21/80	8/80
李特代码（硬）	3/45	3/45	0/45

表1。GPT在学术和专业考试中的表现。在每种情况下，我们模拟真实考试的条件和分数。我们报告了GPT-4的最终分数，根据考试特定的标准进行评分，以及达到GPT-4分数的考生的百分位数。

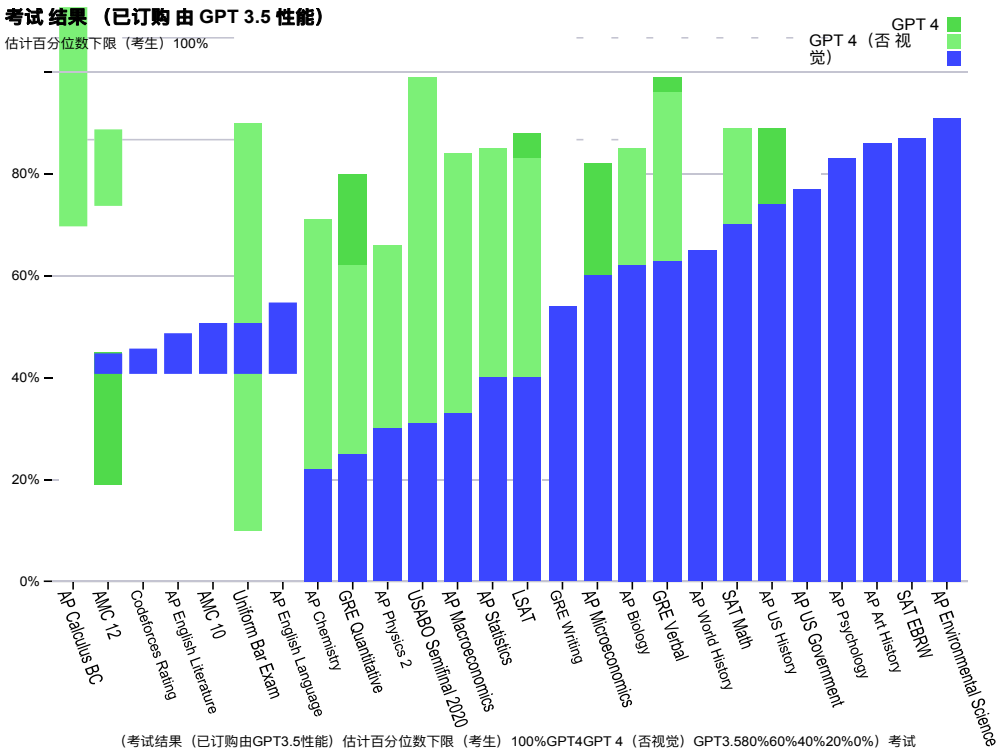


图4。GPT在学术和专业考试中的表现。在每种情况下，我们模拟真实考试的条件和分数。考试根据GPT-3.5的表现从低到高排序。在大多数考试中，GPT 4级优于GPT 3.5级。为了保守起见，我们报告了百分位数范围的低端，但这在AP考试中产生了一些伪像，这些考试有非常宽的评分范围。例如，尽管GPT-4在AP生物学上获得了最高分（5/5），但这在图中只显示为第85个百分位，因为15%的考生获得了该分数。

GPT–4在大多数专业和学术考试中表现出人类水平的表现。值得注意的是，它通过了统一律师考试的模拟版本，分数在考生中排名前10%（表1，图4）。

该模型的考试能力似乎主要源于预训练过程，并没有受到RLHF的显著影响。在多项选择题上，基础GPT–4模型和RLHF模型在我们测试的考试中平均表现相同（见附录B）。

我们还在为评估语言模型而设计的传统基准上评估了预训练的基本GPT–4模型。对于我们报告的每个基准测试，我们对出现在训练集中的测试数据进行污染检查（关于每个基准测试污染的详细信息，请参见附录D）。4在评估GPT–4.5时，我们对所有基准测试都使用了少量提示[1

GPT–4大大优于现有的语言模型，以及以前最先进的（SOTA）系统，这些系统通常具有特定于基准的工艺或额外的训练协议（表2）。

许多现有的ML基准都是用英语编写的。为了初步了解GPT–4在其他语言中的功能，我们使用

Azure Translate将MMLU基准【29，30】（一套跨越57个主题的多项选择题）翻译成多种语言（参见附录F中的翻译和提示示例）。我们发现GPT-4优于GPT 3.5和现有语言模型（Chinchilla[2]和PaLM[3]）的英语语言性能

4在我们的污染检查中，我们发现BIG-bench[42]的某些部分无意中混入了训练集，因此我们将其从报告的结果中排除。
5对于GSM-8 K，我们在GPT-4的预训练组合中包含了部分训练集（详见附录E）。

我们在评估时使用思维链提示【11】。