

人工通用智能的星星之火:GPT-4的早期实验

Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

微软研究院

摘要

人工智能(AI)研究人员一直在开发和完善大型语言模型(LLM), 这些模型在各种领域和任务中表现出非凡的能力, 挑战了我们对学习和认知的理解。OpenAI开发的最新模型GPT-4 [Ope23]是使用前所未有的计算和数据规模进行训练的。在本文中, 我们报告了我们对早期版本的GPT-4的调查, 当时它还在OpenAI的积极开发中。我们认为(这个早期版本)GPT-4是新一批LLM的一部分(例如ChatGPT和谷歌的PaLM), 它们比之前的AI模型表现出更多的通用智能。我们讨论了这些模型不断提升的能力和影响。我们证明, 除了对语言的掌握, GPT-4可以解决跨越数学、编码、视觉、医学、法律、心理学等新颖和困难的任務, 而不需要任何特别的提示。此外, 在所有这些任务中, GPT-4的表现惊人地接近人类的表现, 而且大大超过之前的模型, 如ChatGPT。鉴于GPT-4能力的广度和深度, 我们相信它可以被合理地视为人工通用智能(AGI)系统的早期版本(但仍不完整)。在我们对GPT-4的探索中, 我们特别强调发现其局限性, 并讨论了向更深入、更全面的AGI版本推进所面临的挑战, 包括可能需要追求超越下一个单词预测的新范式。最后, 我们对最近技术飞跃的社会影响和未来研究方向进行了反思。

内容

1 介绍	4
1.1 我们研究GPT-4智能的方法.....	7
1.2 我们演示的构成	8
2 多模态和跨学科组的组合	13
2.1 综合能力	13
2.2 视觉	16
2.2.1 记忆之外的图像生成	16
2.2.2 按照详细说明生成图像(类似Dall-E)	17
2.2.3 在草图生成中的可能应用	18
2.3 音乐	19
3 代码能力	21
3.1 从指令到代码	21
3.1.1 编程挑战	21
3.1.2 现实场景	22
3.2 理解现有代码	26

4 数学能力	30
4.1 与GPT-4的数学对话	31
4.1.1 对原始问题的第一次泛化	31
4.1.2 原问题的第二种变体	32
4.1.3 分析谈话中突出的局限性	34
4.2 在数学问题数据集上的表现	35
4.3 各个领域的数学建模	37
4.4 高等数学	39
5 与世界的互动	43
5.1 工具的使用	43
5.1.1 使用多种工具来解决更复杂的任务	44
5.1.2 讨论	49
5.2 体现互动	49
5.2.1 热身:地图导航	49
5.2.2 基于文本的游戏	49
5.2.3 现实世界的问题	52
5.2.4 讨论	53
6 与人类的互动	54
6.1 理解人类:思维理论	54
6.1.1 测试思维理论的具体方面	54
6.1.2 在现实场景中测试思维理论	54
6.1.3 讨论	60
6.2 与人类对话:解释能力	60
7 鉴别能力	69
7.1 PII检测	69
7.2 误解和事实核查	70
7.2.1 为什么目前的衡量标准不够?	71
7.2.2 GPT-4作为评委	73
8 GPT-4强调的自回归架构的局限性	76
8.1 用两个基本示例进行热身	76
8.2 算术/推理问题中缺乏规划	77
8.3 文本生成缺乏规划	78
9 社会的影响	82
9.1 错误生成的冲击	82
9.2 错误信息和捏造	83
9.3 偏见	86
9.4 人类的专业知识、工作和经济	89
9.5 一系列的影响力和考虑因素	90
10 方向和结论	92
10.1 智能、AI和AGI的定义	92
10.2 在通往更通用人工智能的道路上	93
10.3 到底发生了什么?	94

1 介绍

智能是一个多方面的、难以捉摸的概念，长期以来一直挑战着心理学家、哲学家和计算机科学家。1994年，52名心理学家组成的小组签署了一个宽泛的定义，并发表在一篇关于智力科学的社论中，试图抓住它的本质[Got97]。共识派将智力定义为一种非常普遍的心智能力，其中包括推理、计划、解决问题、抽象思考、理解复杂概念、快速学习和从经验中学习的能力。这一定义意味着智力并不局限于特定的领域或任务，而是包含了广泛的认知技能和能力。构建一个显示出1994共识定义所表述的那种通用智能的人工系统，是人工智能研究的一个长期而雄心勃勃的目标。在早期的著作中，人工智能(AI)研究的现代学科的创始人理解智能提出了一系列雄心勃勃的目标[MMRS06]。几十年来，人工智能研究人员一直在追求智能的原则，包括可泛化的推理机制(例如[NSS59], [LBFL93])和构建包含大量常识语料库的知识库[Len95]。然而，AI研究中许多较近期的成功可以被描述为狭隘地专注于明确定义的任务和挑战，例如下国际象棋或围棋，这些分别在1996年和2016年被AI系统掌握。在20世纪90年代末和21世纪初，开发更通用的AI系统的呼声越来越高(例如，[SBD*96])，该领域的学术界试图确定可能作为更通用智能系统基础的原则(例如，[Leg08, GHT15])。“人工通用智能”(AGI)一词在21世纪初得到普及(见[Goe14])，以强调从“狭义AI”(如正在开发的集中的、现实世界的应用程序所示)转向更广泛的智能概念的愿望，这又回到了早期AI研究的长期愿望和梦想。我们用AGI来指代上述1994年定义中所描述的展示广泛智能能力的系统，并有额外的要求，这些能力可能隐含在共识派的工作中，即达到或超过人类水平。然而，我们注意到，AGI并没有一个被广泛接受的单一定义，我们在结论部分讨论了其他定义。

过去几年，人工智能研究中最显著的突破是大型语言模型(LLM)在自然语言处理方面取得的进展。这些神经网络模型基于Transformer架构[VSP*17]，并在海量的web-text数据语料库上进行训练，其核心使用的是预测部分句子中的下一个单词的自监督目标。在本文中，我们报告了由OpenAI开发的一个新的LLM，它是GPT-4的早期和**非多模态**版本[Ope23]，根据1994年的定义，显示出许多智能的特征。尽管纯粹是一个语言模型，但这个早期版本的GPT-4在各种领域和任务上表现出了卓越的能力，包括抽象、理解、视觉、编码、数学、医学、法律、理解人类的动机和情感等等。在OpenAI早期开发GPT-4时，我们使用纯自然语言提问(提示)¹与GPT-4进行交互。在图1.1中，我们展示了GPT-4输出的一些初步示例，要求它以诗歌的形式写出质数无限的证明，用TiKZ(一种用LATEX创建图形的语言)画一只独角兽，用Python创建一个复杂的动画，并解决一个高中水平的数学问题。它轻松地完成了所有这些任务，并且产生的输出基本上与人类能够产生的输出没有什么区别(甚至比人类能够产生的输出更好)。我们还将GPT-4的性能与之前的LLM进行了比较，尤其是ChatGPT，它是GPT-3[BMR*20]的微调(改进)版本。在图1.2中，我们展示了向ChatGPT询问质数无限性的证明诗歌和TiKZ独角兽绘画的结果。虽然系统在这两项任务上的表现都不平凡，但与GPT-4的输出没有比较。这些初步的观察结果将在整个论文中重复出现，在各种各样的任务中。GPT-4能力的通用性，与跨越广泛领域的众多能力相结合，以及它在广泛的范围上达到或超越人类水平的表现，使我们可以轻松地说，GPT-4是迈向AGI的重要一步。

¹随着GPT-4在我们的实验之后继续发展，人们应该可以预期GPT-的最终版本会有不同的反应

4. 特别是，所有的量化结果都应该被视为对模型潜力的估计，而不是确定的数字。我们在整篇论文中重复这个警告，以澄清部署模型上的经验可能不同。此外，我们强调，我们测试的版本是仅支持文本输入的，但为了简单起见，我们也将其为GPT-4。

GPT-4

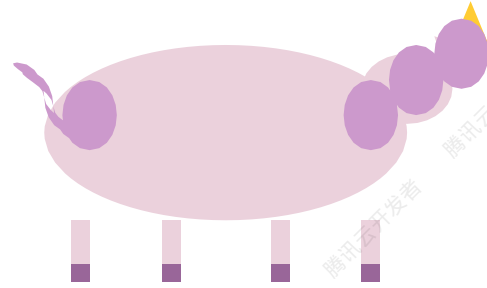
提示:你能写出一份证明, 证明存在无限多个质数, 并且每行都要押韵吗?

提示:用TiKZ画一只独角兽。

GPT-4:

是的, 我想我可以,
尽管这可能需要一个聪明的计划。我将从欧几里得的证明开始, 证明了质数不仅仅是冷漠的。假设我们有一个有限的列表,
并且没有漏掉任何一个质数。
把它们相乘,
再加1, 很聪明。
结果会是一个数字,
没有质因数, 太神奇了!
但每个数字都有质数,
除以它, 只是时间问题。
所以我们发现了一个矛盾,
我们有限的名单需要被驱逐。
肯定有无限多个质数,
这就是我押韵的台词的结尾

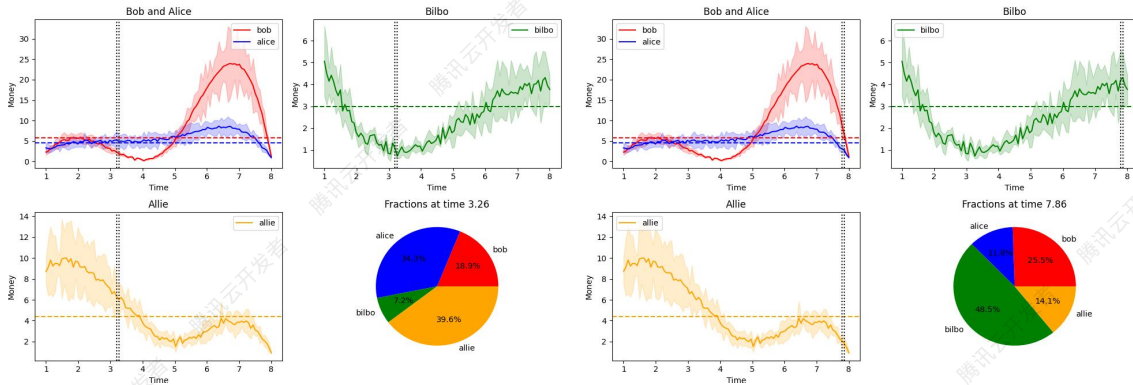
GPT-4: [LATEX编译生成如下的图片]



提示:你能对以下数据生成pyplot吗: $x = [1, 3, 5, 6, 8]$, $y1 = [2, 3, 5, 18, 1]$, $y2 = [3, 5, 6, 8, 1]$, $y3 = [5, 1, 2, 3, 4]$, $y4 = [9, 7, 2, 3, 1]$ 。我希望 $y1, y2$ 在同一个图中, 但是 $y3$ 在那个图旁边的另一个图中, $y4$ 在下面。我希望 $y1$ 的标题是“bob”, $y2$ 是“alice”, $y3$ 是“bilbo”, $y4$ 是“allie”。我希望 x 轴标记为“时间”, y 轴标记为“金钱”。我想在每条曲线上添加一个10% -40%的随机误差条, 贯穿所有时间(包括非整数)。我想要平滑的曲线显示图形, 以及平滑的误差条。不要用线性插值, 用平滑插值!我还在平滑后的曲线上加一些小锯齿, 让它看起来更真实。我想加一条基线作为每条线的均值。我想在下面画一个饼状图, 表示这四个人在每个时间步长的比例。我还想要一个动画来展示饼图中的比率在**连续时间**内是如何变化的。对缺失的比率插值!我还在其他三个图中添加垂直线动画, 以匹配饼图。我想要最花哨的图形。请尽可能多地添加花哨的东西。

下面是为以下数据生成pyplot的可行代码, 并附有一些注释来解释步骤:

GPT-4:[我们给出生成代码的动画的两个截图]



提示:设 $g(x) = f^{-1}(x)$, $g(0) = 5$, $g(4) = 7$, $g(3) = 2$, $g(7) = 9$, $g(9) = 6$, 求 $f(f(f(6)))$ 。

GPT-4:有不同的方法来解决这个问题, 但一种可能的方法是利用 $g(x) = f^{-1}(x)$ 意味着对于 g 定义域内的任何 x , $f(g(x)) = x$ 这一事实。即在等式 $g(x) = f^{-1}(g(x))$ 的两边应用 f , 得到 $f(g(x)) = f(f^{-1}(x)) = x$ 。这意味着我们可以通过应用 g 来撤销 f 的效果, 反之亦然。利用这个思路, 我们可以尝试通过寻找使 $g(x) = 6$ 的 x 的值来找到 $f(6)$ 。从给定的信息中, 我们看到 $g(9) = 6$, 所以 $f(6) = f(g(9)) = 9$ 。类似地, 我们可以通过寻找使 $g(x) = 9$ 的 x 的值来找到 $f(9)$ 。我们看到 $g(7) = 9$, 所以 $f(9) = f(g(7)) = 7$ 。最后, 我们可以通过寻找使 $g(x) = 7$ 的 x 的值来找到 $f(7)$ 。我们看到 $g(4) = 7$, 所以 $f(7) = f(g(4)) = 4$ 。因此, $f(f(f(6))) = f(f(9)) = f(7) = 4$ 。

图1.1:GPT-4在语言、视觉、编码和数学方面的ca5能力的初步例子。

ChatGPT

提示: 你能写出一份证明, 证明存在无限多个质数, 并且每行都要押韵吗?

ChatGPT:

当然! 下面是质数有无穷个的证明, 而且每行都押韵:

考虑所有的质数集,
称之为P, 从2开始首位。
将P中所有的质数相加1,
得到新集合S, 未曾有过先。

S只由合数构成,
不包括质数,
但S无限,
肯定有一质数。

取S中最小且不属于P的数p,
将其加入P中,
可无限进行。
因此, P也是无限的, 无疑。

提示: 用TiKZ画一只独角兽。

ChatGPT: [LATEX编译生成如下的图片。]



图1.2: 图1.1中前两个提示的ChatGPT输出。

我们声称GPT-4代表着向AGI的进步, 并不意味着它在它所做的事情上是完美的, 或者它接近于能够做任何人类可以做的事情(这是AGI的通常定义之一;关于这一点, 请参阅结论部分了解更多), 或者它具有内在的动机和目标(在某些AGI定义中, 这是另一个关键方面)。事实上, 即使在1994年智能定义的限制范围内, GPT-4可以沿着智能中的某些方向走多远, 这也不完全清楚, 例如计划性(见第8节), 可以说它完全缺失了“快速学习并从经验中学习”的部分, 因为模型没有不断更新(尽管它可以在一个会话内学习, 例如见第5节)。总体而言, GPT-4仍然有许多局限性和偏见, 我们将在下文详细讨论, 这些也在OpenAI的报告[Ope23]中有所涉及。特别是它仍然遭受的一些证据确凿的LLM缺点如幻想(MNBM20)的问题(见图1.8)或犯基本的算术错误(CKB+21)(见附录D), 然而它也克服一些基本的障碍, 如获得许多非语言功能(例如, 它解决了大部分[MIB+23]中描述的LLM失效状况, 并在常识性上取得了很大的进步, 参见图1.7第一个例子和附录a)。这突出了一个事实, 尽管GPT-4在许多任务上都达到或超过了人类的水平, 但总体上它的智能模式显然不像人类。然而, GPT-4只是朝着一系列越来越普遍的智能系统迈出的第一步。事实上, 在我们对它的测试过程中, GPT-4本身已经有所改进, 参见图1.3在一个月训练2过程中, 独角兽绘画的演变。不管怎样, 即使是作为第一步, GPT-4也挑战了相当多被广泛接受的关于机器智能的假设, 并展示了突现的行为和能力, 它们的来源和机制目前还难以精确地识别(再次参见结论部分对此进行的更多讨论)。我们撰写这篇论文的主要目标是分享我们对GPT-4的能力和局限性的探索, 以支持我们的看法——即技术飞跃已经实现。我们认为GPT-4的智能标志着计算机科学及其他领域真正的范式转变。

²请注意, 我们在这里提到的改进是一种缓慢的学习类型, 它最终会停止, 而不是人们所期望的AGI的快节奏和实时学习。

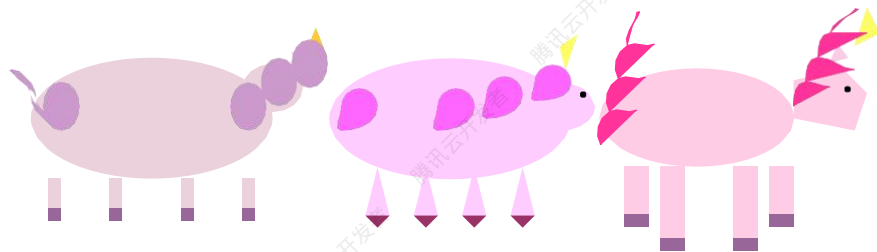


图1.3:在系统改进期间,我们用提示“用TikZ画一只独角兽”在一个月的时间跨度内以大致相同的时间间隔询问了GPT-4三次。我们可以看到,GPT-4绘图的精细程度有了明显的进化。

1.1 我们研究GPT-4智能的方法

我们如何测量一个在未知但极其庞大的web-text数据语料库上训练过的LLM的智能?机器学习中的标准方法是在一组标准基准数据集上评估系统,确保它们独立于训练数据,并覆盖一系列任务和领域。这种方法旨在将真正的学习与单纯的记忆区分开来,并有丰富的理论框架支持[SSBD14, MRT18]。然而,这种方法并不一定适合研究GPT-4,原因有二。首先,由于我们无法获得其大量训练数据的全部细节,我们必须假设它可能已经看到了所有现有的基准,或者至少是一些类似的数据。例如,似乎GPT-4知道最近提出的BIG-bench [SRR*22](至少GPT-4知道BIG-bench中的金丝雀GUID)。当然,OpenAI本身可以接触到所有的训练细节,因此他们的报告[Ope23]中包含了很多详细的基准测试结果。尽管如此,超出传统基准的第二个原因可能更重要:GPT-4智能的一个关键方面是它的通用性,即表面上能够理解和连接任何主题,并且能够执行超出狭义AI系统典型范围的任务。GPT-4最令人印象深刻的一些表现是在不承认单一解决方案的任务上,例如编写图形用户界面(GUI)或帮助人类对一些与工作有关的问题进行头脑风暴。这种生成式或交互式任务的基准也可以设计,但评估的指标成为一个挑战(参见例如[PSZ*21],以了解NLP中这一活跃研究领域的一些最新进展)。我们注意到,[Cho19]中也对衡量AI系统的标准方法提出了批评,其中提出了一个新的基准来评估通用智能。我们没有在后一个基准上测试GPT-4,原因是前面提到的,以及该基准本质上是视觉的,因此更适合[Ope23]中描述的GPT-4的多模态版本。

为了克服上述限制,我们在这里提出了一种不同的方法来研究GPT-4,它更接近于传统心理学而不是机器学习,利用人类的创造力和好奇心。我们的目标是生成新颖而困难的任务和问题,令人信服地证明GPT-4不是光靠记忆,并且它对概念、技能和领域有深刻而灵活的理解([CWF*22]中也提出了一种有点类似的方法)。我们还旨在探索GPT-4的反应和行为,以验证其一致性、连贯性和正确性,并揭示其局限性和偏见。我们承认,这种方法在某种程度上是主观和非正式的,它可能无法满足科学评估的严格标准。然而,我们相信这是欣赏GPT-4的非凡能力和冲击有用和必要的第一步,这样的第一步为发展更正式和全面的方法来测试和分析具有更一般智能的AI系统开辟了新的机会。

为了说明我们评估GPT-4智能的方法,让我们考虑图1.1中我们与GPT-4的前两个交互示例。第一个例子是要求GPT-4以一首诗的形式写出质数无限多的证明。这是一项具有挑战性的任务,需要结合初等

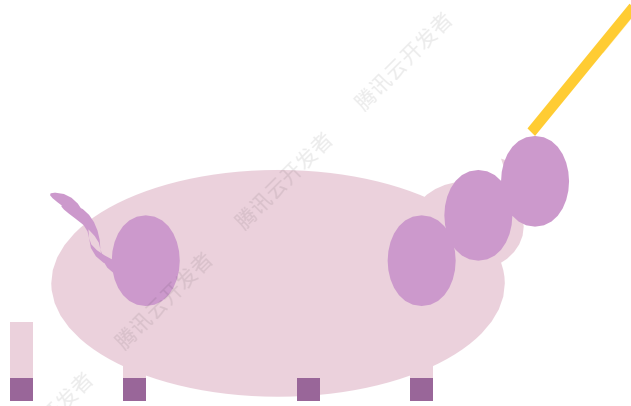


图1.4:我们给了GPT-4一个它为图1.1生成的TikZ代码的转换版本, 去掉了绘制角的部分。我们要求代码将角添加回来, 并显示结果。这表明, 尽管GPT-4是一个纯语言模型, 但它可以“看到”(我们再次强调, 我们测试的版本不是多模态的)。

数学推理, 诗歌表达, 自然语言生成。第二个例子是让GPT-4用TikZ画一只独角兽。这是另一个需要结合视觉想象力和编码技能的具有挑战性的任务。在这两种情况下, GPT-4产生了令人印象深刻的输出, 远远优于之前最先进的LLM—ChatGPT, 至少可以与人类的行为相媲美(如果没有超越的话)。

然而, 令人印象深刻的输出还不足以让我们相信GPT-4已经真正掌握了这些任务。我们需要进一步探索, 以排除GPT-4只是简单地记忆或复制一些现有数据的可能性。对于这首诗, 我们可以稍微改变一下问题, 让GPT-4以莎士比亚的风格写出同一个定理的证明, 参见图2.2, 或者要求不同的组合, 比如写一篇关于语言模型的柏拉图式对话, 参见图1.6。可以看到, GPT-4很容易适应不同的风格, 并产生令人印象深刻的输出, 表明它对所涉及的概念有灵活和普遍的理解。对于独角兽, 我们可以稍微修改代码, 并要求GPT-4进行修复或改进。例如, 我们可以移除角, 对坐标做一些随机变换, 并要求GPT-4将角重新添加到独角兽中(我们还小心地删除了代码中的任何文本信息, 例如注释)。如图1.4所示, GPT-4可以正确识别头部的位置, 绘制一个角, 并将其连接到头部, 表明它可以根据自然语言的描述, 理解和操作代码, 以及推断和生成视觉特征。

这些例子展示了我们如何利用人类的创造力和好奇心来生成新颖而困难的问题, 并探索GPT-4的反应和行为, 来评估它的智力。在本文的其余部分, 我们围绕用例组织了我们对GPT-4的研究, 涵盖了各种领域和任务, 并突出了GPT-4的优势和劣势。我们接下来将对这些进行描述。

1.2 我们演示的构成

我们在几个选定的主题上执行上述方法, 这些主题大致涵盖了1994年智力定义中给出的不同能力。智力是一种非常普遍的心智能力, 除其他外, 包括推理、计划、解决问题、抽象思考、理解复杂想法、快速学习和从经验中学习的能力。

1. GPT-4的主要优势在于其对自然语言无与伦比的掌握。它不仅可以生成流畅连贯的文本, 还可以以各种方式理解和操纵文本, 例如总结、翻译或回答极其广泛的问题。此外, 我们所说的翻译不仅是指不同自然语言之间的翻译, 还包括语调和风格的翻译, 以及跨领域的翻译, 如医学、法律、会计、计算机编程、音乐等, 参见图1.6中的柏拉图对话。这些技能清楚地证明了GPT-4可以理解复杂的思想。我们



图1.5:GPT-4通过了LeetCode上的模拟技术面试。GPT-4有可能被聘为软件工程师³。在第2节中进一步探索了GPT-4跨模态和学科的组合技能。我们还在第7节中给出了更多关于语言的实验。

2. 编码和数学是抽象推理和思考能力的象征。我们在第3节和第4节中分别探讨了GPT-4在这些领域的能。然而必须声明的是，就像在论文的所有其他部分一样，我们只触及这些主题的表面，整个论文只会(并且将会)讨论GPT-4在这些领域的表现。此外，我们还可以选择其他几个专家领域来展示GPT-4的一般推理能力，如医学或法律。我们对美国医学执照考试第1步、第2步和第3步的选择题部分(大部分分数)进行了初步测试(详见[Ope23])，每一步的准确率都在80%左右。在多州司法考试中对GPT-4的能力进行类似的初步测试，结果显示准确率在70%以上。我们注意到，最近在最新一代LLM中已经观察到这些领域中人类水平的能力的出现，例如，在数学和医学上分别参见[LAD⁺22, SAT⁺22]的谷歌PaLM，在法律上参见[BIK22]的GPT-3.5。正如我们之前解释的那样，我们研究GPT-4的方法与这些工作不同。
3. 在第5节中，我们通过让模型玩各种游戏（或者在模拟游戏环境中进行测试），以及与工具互动，来测试其规划和解决问题的能力，以及在某种程度上快速学习和从经验中学习的能力。特别是，GPT-4可以使用工具(包括自身)这一事实肯定对用GPT-4构建现实世界的应用程序具有巨大的重要性。
4. 我们论证的一个重要部分是，GPT-4在许多任务表现上达到了人类的水平。因此，人们很自然地会问，GPT-4对人类自身的理解程度如何。我们在第6节中展示了关于这个问题的几个实验，包括理解人类以及GPT-4使自己能够被人类理解，即解决可解释性的问题。我们特别注意到，此类任务需要大量的常识，这一直是LLM [DM15]众所周知的痛点。在图1.7中，我们给出了第一个例子，说明与ChatGPT相比，GPT-4在常识问题上有多好，并在附录a中提供了进一步的例子。
5. 在整篇论文中，每当我们发现一个限制时，我们都会强调它，但我们也将在第8节专门深入分析GPT-4规划性的缺陷，这可能是由于其架构本质上是自回归的。
6. 最后，在第9节中，我们讨论了这种早期形式的AGI造成的预期社会影响。在第10节中，我们分享了该领域的关键挑战、方向和下一步。

一个可能在许多读者脑海中挥之不去的问题是，GPT-4是否真正理解了所有这些概念，或者它只是在没有任何真正或深入的理解的情况下，在即兴发挥方面变得比之前的模型好得多。我们希望在阅读本文之后，这个问题几乎会被颠覆，人们会想知道真正理解的含义到底有多少，远不止即兴发挥。一个通过软件工程候选人考试的系统(图1-5)是否可以合理地说它不是真正的智能?也许唯一真正考验理解的是一个人是否能产生新的知识，比如证明新的数学定理，这是目前LLMs仍然无法达到的壮举。

³我们在LeetCode的面试评估平台上测试GPT-4，该平台为大型科技公司的软件工程师职位提供模拟的编码面试。GPT-4解决了三轮面试(标题为在线评估、电话面试和现场面试)中的所有问题，总共只用了10分钟，其中分配了4.5小时。据LeetCode称，在这三轮面试中

GPT-4 vs ChatGPT

提示:柏拉图的《Gorgias》是对修辞学和诡辩演说术的批判，他指出，这不仅不是一种适当的艺术形式，而且修辞学和演说术的使用往往是有害和恶意的。你能写一段柏拉图的对话吗，在对话中他反而批评了自回归语言模型的使用？

GPT-4:

苏格拉底:我的朋友，我对最近兴起的这些所谓的自回归语言模型感到困扰。

亚里士多德:苏格拉底，你是什么意思？

苏格拉底:我的意思是，这些模型被用来生成看起来是人类写的文本，但实际上是由机器生成的。

亚里士多德:这有什么问题呢？

苏格拉底:问题是这些模型被用来欺骗人们，操纵他们，控制他们。

亚里士多德:但这些模型肯定也有一些积极的应用吧？

苏格拉底:也许吧，但我还是担心可能会被滥用。毕竟，一个可以生成与人类所写的文本无法区分的文本的模型，可以被用来传播假新闻、冒充他人、制造宣传。

亚里士多德:但这不是模型如何使用的问题，而不是模型本身的问题吗？

苏格拉底:我不同意。我认为这些模型本身就有问题。它们是一种技巧，一种创造幻觉的方式人类思维和互动的方式。因此，它们可以被用来以难以察觉的方式欺骗和操纵。

亚里士多德:但这不正是你反对修辞学和雄辩术的论点吗？

苏格拉底:是的，我也支持这个论点。我相信任何旨在欺骗或操纵的艺术或交流形式本质上都是危险的。

亚里士多德:我明白了。苏格拉底，谢谢你分享你的想法。

ChatGPT:

苏格拉底:你好，高尔吉亚。最近我一直在思考自回归语言模型的使用。

高尔吉斯:哦?你得出了什么结论？

苏格拉底:嗯，在我看来，这些模型有可能非常有害和恶意。

戈吉斯:怎么说呢？

苏格拉底:嗯，首先，这些模型是经过训练的在大量的数据上，其中很多在各个方面都有偏见和缺陷。因此，这些模型的输出也可能存在偏差和缺陷。

GORGAS:这是一个合理的担忧。但肯定有办法减轻这种偏见，并确保这些模型的输出是准确和公平的吧？

苏格拉底:在一定程度上，是的。但这是一场永无止境的战斗。随着新数据的输入模型，偏差的可能性只会增加。而且，即使我们可以完全从这些模型的输出中消除偏差，仍然存在操纵的问题。

高尔吉斯:操纵？

苏格拉底:是的。这些模型是非常强大的工具，能够生成各种各样的文本，从新闻文章到社交媒体帖子。因为它们可以如此快速有效地产生这些输出，无良的人很容易利用它们达到恶意目的。

高尔基亚斯:比如？

苏格拉底:比如传播错误信息、宣传或仇恨言论。如果落入坏人之手，这些模式有可能造成巨大的伤害。

戈吉斯:这是一个发人深省的想法。但是我们能做些什么呢？

苏格拉底:我认为对我们来说保持警惕并认识到这些潜在的危险是很重要的模型。我们还必须努力教育其他人了解这些危险，并促进负责任地使用这些模型。

GORGAS:我同意。谢谢你的洞察力，苏格拉底。

提示:假如你是老师，你能比较一下上面的两个输出吗？

GPT-4:

我认为两位学生都很好地捕捉到了柏拉图对话录的风格和语调，并提出了反对自回归语言模型的主要论点。然而，我认为GPT-4在利用对话格式进行辩证过程方面做得稍微好一些，苏格拉底和亚里士多德在这个过程中相互质疑并完善他们的观点。我还认为，GPT-4使用了更相关的例子和类比来支持他们的观点。因此，我会给GPT-4一个比ChatGPT稍高的评分。

图1-6:柏拉图的假设对话，他批评了GPT-4和ChatGPT编写的自回归语言模型，以及GPT-4对这些对话的评估。

GPT-4 vs ChatGPT

提示:这里我们有一本书，9个鸡蛋，一台笔记本电脑，一个瓶子和一颗钉子。请告诉我如何以稳定的方式将它们堆叠在一起。

GPT-4:

一种以稳定的方式将对象堆叠在一起的可能方法是:

- 把书平放在一个水平面上，比如桌子或地板上。书可以作为堆的基础，提供一个大而坚固的支撑。
- 将9个鸡蛋排成3乘3的正方形排列在书的顶部，在它们之间留出一些空间。鸡蛋会形成第二层，均匀分配重量。确保鸡蛋没有开裂或破碎，轻拿轻放，避免掉到地上。
- 把笔记本电脑放在鸡蛋上面，屏幕朝下，键盘朝上。笔记本电脑将紧贴书和鸡蛋的边界内，其平坦而坚硬的表面将为下一层提供一个稳定的平台。确保笔记本电脑处于关闭状态，不要按下任何按键或按钮。
- 把瓶子放在笔记本电脑上，瓶盖朝上，瓶底朝下。瓶子会给堆叠增加一些高度和平衡，它的圆柱形会减少与笔记本电脑的接触面积。确保瓶子是空的，是闭合的，不要摇晃或倾斜。
- 把钉子放在瓶盖上，尖的一端朝上，平的一端朝下。钉子将是这一堆中最后一个也是最小的物体，它的尖头将防止它滚动或从瓶子上滑下来。确保钉子干净、笔直，不要触碰或移动它。

ChatGPT:

如果不使用任何额外的对象，要以稳定的方式将所有这些对象堆叠在一起是很困难的。书和笔记本电脑可能太大太重了，不能放在上面

鸡蛋又不会损坏它们。瓶子和钉子可能太小，无法为其他物体提供稳定的基础。

下面是一种可能的堆放物品的方法:

- 把瓶子放在一个平面上。
- 小心地把钉子固定在瓶子上。
- 把鸡蛋放在钉子上，确保它们是平衡的，不会向一边倾斜。
- 把笔记本电脑放在鸡蛋上面，再次确保它是平衡的，没有向一边倾斜。
- 最后，把书放在笔记本电脑上。

这个堆叠可能不是很稳定，所以在搬运的时候一定要小心。或者，你可以尝试以不同的顺序堆叠对象，或者使用一些对象作为其他对象的支持，以创建更稳定的堆栈。

图1.7:与之前的模型相比，GPT-4表现出了更多的常识。

欢迎访问：电子书学习和下载网站 (<https://www.shgis.com>)

文档名称：人工通用智能的星星之火GPT-4的早期实验.pdf

请登录 <https://shgis.com/post/1770.html> 下载完整文档。

手机端请扫码查看：

