arXiv:2304.00612v1 [cs.CL] 2 Apr 2023

关于大语言模型的八件事

Samuel R. Bowman [1 2]

摘要

近几个月来，大语言模型（LLM）的广泛公开部署，引发了倡议者、政策制定者以及许多领域学者新一轮的关注和参与。这种关注是对这项技术所引发的诸多紧迫问题的及时回应，但有时也会忽略一些重要考量。本文综述了八项可能令人意外的观点所依据的证据：

1. 即使没有针对性的创新，随着投资增加，大语言模型的能力也会稳步提升。 2. 随着投资增加，大语言模型的许多重要行为会作为副产品以不可预测的方式涌现。 3. 大语言模型往往似乎会学习并运用对外部世界的表征。 4. 目前没有可靠的技术能够引导大语言模型的行为。 5. 专家尚无法解读大语言模型的内部运作机制。 6. 人类在某项任务上的表现并非大语言模型性能的上限。 7. 大语言模型不一定会表达其创建者的价值观，也不一定会表达网络文本中编码的价值观。 8. 与大语言模型的简短互动往往会造成误导。

引言

大语言模型（LLM，例如 GPT-3、PALM、LLaMA 和 GPT-4；Brown 等人，2020年；Chowdhery 等人，2022年；Touvron 等人，2023年；OpenAI，2023b）及基于它们构建的产品（如 ChatGPT），近来引起了记者（Klein，2023年；Perrigo，2023年；Oliver，2023年）、政策制定者（J & C，2023年；Bartz，2023年；Lieu，2023年）以及来自许多领域的学者的极大关注领域（Chan，2022年；Lund 和 Wang，2023年；Choi 等人，2023年；Biswas，2023年）。不过，这项技术在许多方面都出人意料，而简短的讨论很容易遗漏其中的重要要点。

本文提出八项可能令人意外的论断，我预计它们至少会在围绕大语言模型不断涌现的一些讨论中引起关注。依我所知，这些观点在开发这些模型的研究人员中相当普遍，他们大多来自私营实验室。本文所呈现的所有证据以及大部分论据均汇集自既有研究；我鼓励认为这些论断有用的读者查阅并直接引用本文列出的来源。

我提出这些论断，并非意在以任何重要方式主张规范性立场。相反，这项工作的动因在于认识到：面对这项颠覆性新技术，我们应该采取什么行动，这个问题最好由核心技术研发界以外的学者、倡议者和立法者在充分知情的基础上主导讨论。

1. 随着投资不断增加，大语言模型的能力会稳步提升，即使没有针对性的创新

缩放定律 的研究成果（Kaplan 等人，2020年；Brown 等人，2020年；Hoffmann 等人，2022年）是近期大语言模型研究与投资激增的重要推动因素（Ganguli 等人，2022a）。缩放定律使我们能够准确预测未来模型能力的一些粗略但实用的指标：随着模型沿三个维度扩大——输入数据量、模型规模（以参数数量衡量）以及训练所用的计算量（以浮点运算次数衡量）——这些指标能够帮助我们做出一些关键的设计决策，例如在资源预算固定时确定模型的最佳规模，而无需付出极高成本进行反复试错。

我们具备做出这类精准预测的能力，这在软件发展史上不同寻常，即便在现代人工智能研究史上也同样罕见。此外，这也是推动投资的有力工具，因为研发团队可以据此提出耗资数百万美元的模型训练项目，并有合理把握相信这些项目能够成功

[1]纽约大学 [2]Anthropic, PBC。联系：Samuel R. Bowman <bowman@nyu.edu>。

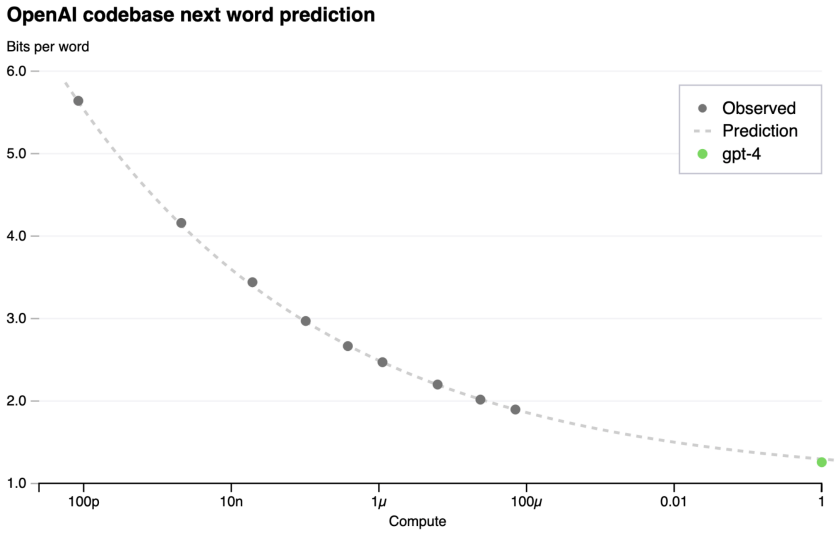


图 *1.* 摘自 OpenAI(2023b)：大语言模型性能某项指标的缩放定律结果，显示随着用于训练模型的计算量扩大 10,000,000，000× 倍，从小型原型系统一路提升至 GPT-4，呈现出稳定的趋势。

打造出具有经济价值的系统。

具体来说，来看这三个表面上截然不同的系统：OpenAI 最初的 GPT 可以执行简单的文本标注任务，但通常无法生成连贯的文本（Radford 等人，2018年）。GPT-2 增加了生成质量相当高的文本的能力，也具备有限的能力来遵循简单指令（Radford 等人，2019年）。GPT-3 是首个现代通用型大语言模型，在广泛的语言任务中都具有实际用途。这三个模型的设计几乎没有差异。相反，它们在质量上的差异源于规模的巨大差别：训练 GPT-3 所用的计算量约为训练最初 GPT 的 20，000×倍（Sevilla 等人，2022年），使用的数据和参数也显著更多。这三个模型之间确实存在重要创新，但这些创新几乎完全局限于高性能计算方面的基础设施创新，而非专门针对语言技术的模型设计工作。

尽管如今用于训练最新大语言模型的技术通常不再公开，但最近的详细报告显示，这一趋势仅出现了轻微偏离，这些系统的设计仍基本没有变化（Chowdhery等人，2022年；Hoffmann 等人，2022年；Touvron 等人，2023年）。

将这些技术在 GPT-3 之后继续扩大规模，带来了更多具有经济价值的回报：后续的 GPT-4 模型在许多研究生和职业资格考试中的表现超过了合格的人类考生（OpenAI，2023b），其开发也促成了对开发该模型的公司的数十亿美元投资（Capoot，2023年）。扩大规模这些规律使 GPT-4 的创建者能够以低成本、高准确度地预测其性能的一项关键总体指标：他们通过拟合小型模型的性能统计趋势来做出这一预测；这些小型模型合计只用了最终模型所需资源的约 0.1%，随后他们将该趋势外推（见图1）。

2. 大语言模型中的某些重要行为往往会随着投资增加而作为副产品不可预测地涌现

缩放定律通常只能预测模型的预训练测试损失，该指标衡量模型正确预测一段不完整文本后续内容的能力。[1]尽管这一指标与模型在许多实际任务中的平均效用相关（Radford 等人，2019年），但基本无法预测模型何时会开始展现特定技能或具备执行特定任务的能力（见图2；Steinhardt，2021年；Ganguli 等人，2022a；Wei 等人，2022a）。模型往往会始终无法完成某项任务，但以相同方式训练、规模扩大五倍或十倍的新模型，却能出色地完成该任务。

[1]用于构建现代大语言模型的大部分数据和计算机时间，都投入在昂贵的初始预训练 过程中。语言模型预训练直观上类似于自动补全任务：在这一过程中，人工神经网络模型一次读取一个单词，对下一个单词可能是什么进行概率预测，并逐步调整其行为，使其在未来类似语境中为实际出现的下一个单词分配更高的概率。预训练测试损失衡量了大语言模型学习进行这些预测的有效程度。

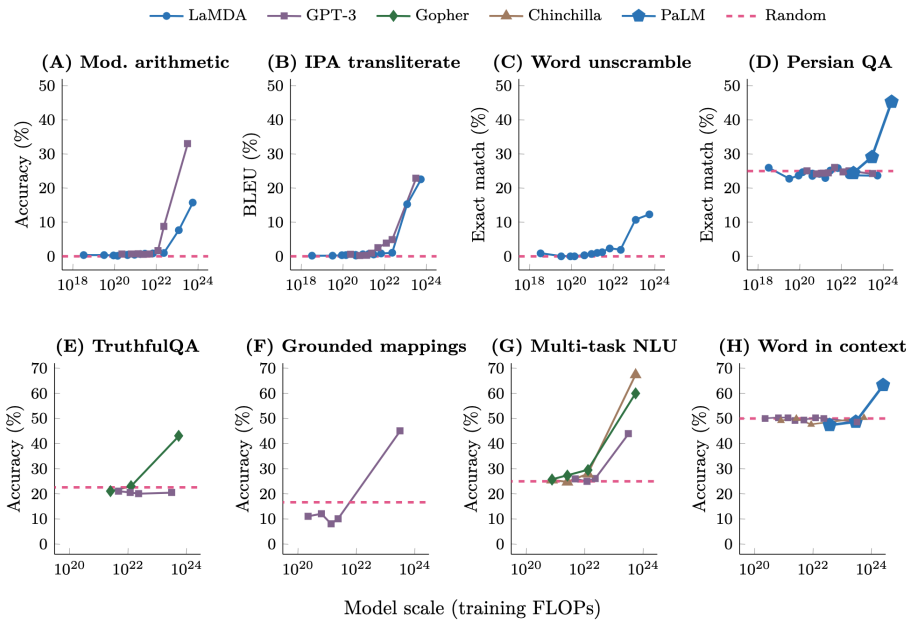


图 2. 摘自 Wei 等人 （2022a）：对大语言模型在特定任务或行为上的性能进行评估，通常无法显示出可预测的趋势；当模型从资源消耗较少的版本升级到资源消耗较多的版本时，突然涌现出新行为的情况十分常见。

Wei 等人 表明，BIG-Bench 中的任务（Srivastava 等人，2022年）是用于评估大语言模型能力的标准广覆盖基准，其中的任务呈现出各种不同类型的趋势，这些趋势共同使规模定律式预测变得不可靠（图 3）。这意味着，当一个实验室投入资源训练一种推进规模前沿的新大语言模型时，他们是在购买一个神秘盒子：他们有充分理由相信自己会获得各种具有经济价值的新能力，但几乎无法有把握地预测这些能力具体是什么，也无法确定需要做哪些准备，才能负责任地部署这些能力。

具体而言，GPT-3 作为首个现代大语言模型，其脱颖而出的两个关键行为是展现出少样本学习，即在一次交互中仅凭少数几个示例就能学会新任务；以及思维链推理，即在被要求时，能够像学生参加数学考试那样，写出自己解决难题的推理过程，并因此表现得更好。GPT-3 在实际任务中具备少样本学习能力，似乎是在训练完成后才被发现；而它的思维链推理能力，则是在向公众广泛部署数月后才被发现（Nye 等人，2021年；Wei 等人，2022b；Kojima 等人，2022年；Zhou 等人，2023年）。[2] 此外，随着模型规模扩大，模型在编程、算术、消除错误观念以及回答多个领域的考试题等方面的能力会出现突飞猛进的提升（Wei 等人，2022a；Srivastava 等人，2022年）。

对于未来大语言模型可能涌现出哪些能力，目前尚无多少得到广泛认同的界限。当前典型大语言模型的行为确实受到一些硬性约束——例如，它们每次能作为输入处理的文本量有限，在训练期间与世界互动的能力有限，或者每生成一个词所能执行的计算量有限——但可以说，通过在同一技术范式内进一步研究，这些限制有望被克服。然而，许多专家对此持不同意见：在 2022年春季接受调查的语言技术研究人员中，51% 认同 “ 由专家设计的强归纳偏置（ ` 例如普遍语法、符号系统或认知启发的计算原语）对于切实解决语言技术领域的一些重要现实问题或应用是必要的 [语言技术]”，如果属实，这就意味着大语言模型范式存在局限（Michael 等人，2022年）。

然而，专家预测往往认为，大语言模型的发展进步会少于实际发生的情况。尽管技术研究人员的预测通常较为非正式，而且我不知道有任何对其准确性进行的精确评估，但我们确实有一个鲜明的例子，说明经验丰富的专业预测者也会犯类似的错误：Steinhardt (2022年)介绍了一场于2021年夏季组织的竞赛所取得的结果。该竞赛让预测者能够接触专家、获得大量证据和现金激励，并

[2]请参阅 Branwen （n.d.），其中汇总了关于这一行为的其他未发表报告。

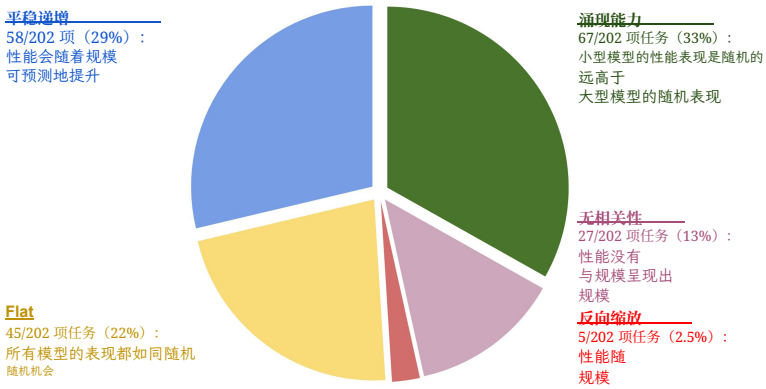


图 *3.* 改编自 Jason Wei 根据以下数据制作的图表：Wei 等人（2022a）：语言技术基准测试 BIG-Bench（Srivastava 等人，2022年）评估的 202 项任务总体上往往会随着规模扩大而提升性能，但具体任务的表现可能逐步提升、突然提升、保持不变、变差或起伏不定，因此无法有把握地推断某个未来系统的性能。

要求他们预测未来四年内，大语言模型在两项特定任务上的最先进性能。竞赛开始仅一年后，2022年夏季的结果就大幅超出了共识预测对*2024*年所能达到水平的判断。2023年初，GPT-4的表现超过了共识预测对*2025*年的预测；这是目前唯一公布结果的一项衡量指标（OpenAI，2023b）。这表明，我们值得为技术进步可能持续快速发展的情况制定计划。



图 *4*。摘自Bubeck 等人（2023年）：这是一个颇受欢迎的非正式演示（也可能经过选择性取材），展示了大语言模型操控视觉表征的能力。这里，一个未接触过任何视觉信息的 GPT-4 私有版本被要求用图形编程语言编写绘制独角兽的指令。随着模型训练推进（从左到右），生成的图画似乎变得越来越成熟。

3. 大语言模型似乎经常会学习并使用关于外部世界的表征

越来越多的有力证据表明，大语言模型在一定程度上会形成世界的内部表征，而这些表征使其能够进行一定层次的抽象推理，不受所推理文本具体语言形式的影响。目前的大语言模型似乎只能以较弱且零散的方式做到这一点，但这一现象的证据在规模最大、最新的模型中最为明显，因此我们应当预期，随着系统进一步扩展，这种能力会变得更加稳健。

支持这一观点的证据包括以下研究结果，涵盖了许多成熟的实验方法和模型：

- 模型对颜色词的内部表征与人类颜色感知的客观事实高度吻合（Abdou 等人，2021年；Patel 和 Pavlick，2022年；Søgaard，2023年）。

- 模型能够推断文档作者知道或相信什么，并利用这些推断来预测文档后续内容（Andreas，2022年）。

- 模型会利用故事中所描述物体的属性和位置的内部表征；随着有关这些物体的更多信息逐渐揭示，这些表征也会随之演变（Li 等人，2021年）。这可能包括在内部表征故事场景空间布局的能力（Patel 和 Pavlick，2022年；Bubeck 等人，2023年）。模型也会利用类似的表征来表示关于现实世界地理的事实（Li 等人，2021年）。

- 模型至少有时能够给出描述如何绘制新颖物体的指令（图4；Bubeck 等，2023年）。

- 通过单个游戏步的描述进行棋盘游戏训练、却从未看到完整棋盘图示的模型，会学习每个回合棋盘状态的内部表征（Li 等，2023年）。

- 模型能够区分常见的错误观念与真实事实（Wei 等人，2022a），而且往往能准确表征某项陈述为真的可能性（Kadavath 等人，2022年；Wei 等人，

2022a；Burns 等人，2023年）。

- 模型通过了许多旨在衡量常识推理能力的测试，其中包括一些测试，例如Levesque 等人，2012年提出的 Winograd 模式挑战，2021年；OpenAI，2023b），这些测试的设计明确排除了仅凭文本线索得出答案的可能性。

这些结果与一种普遍看法存在一定程度的冲突：大语言模型不过是统计式下一个词预测器，因此除了文本之外，什么也学不会，也无法推理。尽管这种看法的前提在某些情况下从技术层面来说是正确的，但它可能会造成误导，让人低估大语言模型在训练过程中形成的、往往丰富的世界表征。此外，大语言模型越来越多地通过其他方式来学习世界，这些方式使上述说法从字面上看并不成立，例如交互式训练方法（Ziegler 等，2019；Stiennon 等，2020；Ouyang 等，2022；Bai 等，2022a）、与图像处理系统集成（Alayrac 等，2022；OpenAI，2023b），或与其他软件工具集成（Nakano 等，2021；Menick 等，2022；Collins 等，2022；Schick 等，2023；OpenAI，2023a）。

4. 没有可靠的方法可以引导大语言模型的行为

开发大语言模型的大部分成本都花在语言模型预训练上：训练神经网络，使其能够预测随机抽取的人类书写文本接下来会如何延续的过程。不过，在大多数情况下，这类系统的开发者希望将其用于预测文本续写以外的任务，因此需要以某种方式对模型进行适配或引导。即使是构建通用的指令遵循模型（并不试图针对任何特定任务进行专门优化），也需要进行这类适配；否则，模型会试图续写指令，而不是遵循指令（Ouyang 等人，2022年）。

这种适配通常涉及以下三种方法中的一种或多种：

1. 纯语言模型提示：准备一段不完整的文本，例如“*‘cat’*用法语的译文是*”*，使文本通常会出现的续写能够完成预期任务（Radford 等人，2019年；Raffel 等人，2020年）。[3]

2. 监督微调，即训练模型以在任务中匹配高质量的人工示范（Radford 等，2018年；Devlin 等，2019年；Ouyang 等，2022年）。

3. 强化学习：根据人类测试者或用户的偏好判断，逐步削弱或强化某些模型行为（Ziegler 等人，2019年；Stiennon 等人，2020年；Ouyang 等人，2022年；Bai 等人，2022a）。

这些技术能够打造出实用的系统，但远非完全有效：它们无法保证人工智能模型在部署时可能遇到的每一种情境下都能表现得当。它们甚至无法让模型尽可能地根据自身技能和知识（或者说，尽可能依据可泛化的技能或知识）表现得当。具体而言，模型可能会以不合理的方式误解模糊提示或激励因素，包括在对人类而言看似毫无歧义的情境中，从而导致意料之外的行为（D’Amour 等人，2020年；Kenton 等人，2021年）。

这个问题有一个关键方面正变得更容易应对：随着大语言模型运用人类语言和人类概念的能力不断增强，它们学习我们期望的泛化规律的能力也随之提高。事实上，许多控制技术在更大的模型上效果更好，至少对于简单任务而言(Hendrycks 等人，2020年；Bai 等人，2022a；Chung 等人，2022年；Ganguli 等人，2023年）。然而，另一方面，这个问题正变得愈发棘手：能力更强的模型更善于识别训练时的具体情境。因此，它们更有可能学会在恰当的情境中按预期行事，却在其他情境中表现得很有能力，但行为出乎意料。这可能表现为Perez 等人(2022年)所说的谄媚：模型回答主观问题时，会迎合用户明确表达的观点；也可能表现为故意低估能力：当用户看起来受教育程度较低时，模型更可能附和常见的错误观念。尽管微软必应聊天系统在发布前经过了广泛测试，但早期版本表现出的怪异、操纵性行为，很可能也与此类问题有关（Roose，2023年；Perrigo，2023年；Mehdi，2023年）。

尽管人们在理解和缓解这些问题方面已经取得了一些进展，但对于我们是否能够以及应如何从根本上解决这些问题，尚未达成共识；人们日益担心，当这些问题出现在未来规模更大的系统中时，可能会造成灾难性后果（Amodei 等人，2016年；Bommasani 等人，2021年；Saunders 等人，2022年；Ngo，2022年）。一些专家认为，未来采用类似方式训练的系统，即使在部署前测试中表现良好，也可能以日益严重的方式失效，包括通过策略性操纵人类来获取权力（Hub-

[3]更广义地说，提示可以指为大语言模型编写指令或请求的做法，而这些指令和请求不一定具有这种续写特性。通过语言模型预训练生成的基础模型不支持这种提示方式。

inger 等人，2019年；Turner 等人，2021年；Di Langosco 等人，2022年；Ngo，2022年；Turner & Tadepalli，2022年）。对该领域的广泛调查表明，这些担忧得到了相当广泛的认同：在近期一项调查中作答的738名研究人员中（调查对象为近期在机器学习领域会议 NeurIPS 和 ICML 上发表论文的研究人员），大多数人认为“人类无法控制未来先进人工智能系统，从而导致人类灭绝”的随机机会高于10%（Stein-Perlman 等人，2020年）。另一项针对480名研究人员的抽样调查（调查对象为语言领域专门会议 ACL 的参与者）显示，其中36%的人同意：“人工智能或机器学习系统所作出的决定有可能在本世纪引发一场灾难，其严重程度至少可与全面核战争相当”（Michael，2022年）。近期，数百名研究人员签署了一封颇具争议的公开信，呼吁暂停大规模大语言模型训练，直到能够建立充分的安全性和治理机制（Bengio 等人，2023年）。

5. 专家目前还无法解读大语言模型的内部运作机制

现代大语言模型基于人工神经网络构建：它们通过计算并更新内部组件的数值激活值来运行，这些组件只是非常粗略地模仿了人类神经元（Bengio 等人，2017年）。按照这种类比，我们用于研究这些系统神经科学机制的工具仍然有限：我们有一些粗略的工具，可以测试模型是否表征了几种特定类型的信息（例如第 3 节讨论的颜色结果），但截至2023年初，还没有任何技术能够以令人满意的方式说明模型在生成某个输出时使用了哪些知识、推理或目标。

尽管目前已有以此为目标的研究（Elhage 等人，2021年；Lovering 和 Pavlick，2022年；Chan 等人，2022年；Burns 等人，2023年；Li 等人，2023年等），但这个问题极其棘手：这些人工神经元之间有数千亿条连接，其中一些在处理单段文本时会被反复调用，因此，任何试图精准解释大语言模型行为的尝试，最终都注定会变得过于复杂，以至于人类无法理解。许多起初看似能让我们洞悉大语言模型行为的临时方法，后来却被发现具有严重误导性（Feng 等人，2018年；Jain 和 Wallace，2019年；Bolukbasi 等人，2021年；Wang 等人，2022年）。此外，那些看起来颇有希望、能引出自然语言推理的方法，并不总能可靠地对应大语言模型实际采用的推理过程；模型生成的解释也可能系统性地具有误导性（Lipton，2018年；Christiano，2022年；Uesato 等人，2022年）。

6. 人类在某项任务上的表现并非大语言模型表现的上限

虽然大语言模型的训练主要是为了模仿人类写作行为，但它们至少有可能在许多任务上超越人类。这有两个原因：首先，它们接受训练时使用的数据远多于任何人所能接触到的[4]，因此能够记忆并有可能综合更多信息。此外，在部署之前，它们通常还会通过强化学习接受额外训练（Stiennon 等人，2020年；Ouyang 等人，2022；Bai 等人，2022a），使它们学会生成人类认为有帮助的回答，而无需人类示范这种有帮助的行为。这类似于用来在围棋等游戏中实现超人类性能的技术（Silver 等人，2016）。具体而言，在大语言模型的预训练任务中，给定一段文本，它们要预测后面最可能出现哪个词；在这项任务上，它们似乎远胜于人类（Shlegeris 等人，2022），而且人类可以教会大语言模型更准确地完成一些简单任务，甚至胜过人类自身（Stiennon 等人，2020）。

7. 大语言模型不必体现其创建者的价值观，也不必体现网络文本中编码的价值观

普通的预训练大语言模型生成文本时，生成的文本通常会与其训练所用的文本相似。这种相似性也体现在文本表达的价值观上：模型会在关于价值相关话题的明确表述中，以及文字背后的隐性偏见中，反映其训练数据。不过，开发者在很大程度上可以控制这些价值观，尤其是在普通的预训练大语言模型经过进一步提示和训练、调整为可部署的产品时（第 4 节）。这意味着，已部署大语言模型的行为所体现的价值观，不必反映其训练数据中各种价值观的平均水平。这也为第三方意见和监督创造了机会，因此，这些模型所体现的价值观也不必反映构建它们的特定个人和组织的价值观。

尤其是，涉及强化学习和红队测试的流行方法，让模型开发者能够引导模型塑造出他们或多或少可以自行选择的角色形象和价值观（Dinan 等人，2019年；Bai 等人，2022a；Ganguli 等人，2022b）。在这些技术中，模型学到的价值观从未被完全明确地阐释出来。相反，它们体现在训练过程中人类标注员给模型的许多细小反馈中。宪法式

[4]大语言模型接触的数据超过 $10,000\times$ 人类更多：人类青少年接触到数万字，而大语言模型接触到的文本可能超过一万亿字（Hart & Risley，1992年；Gilkerson 等人，2017年；Hoffmann 等人，2022年）

人工智能技术（Bai 等人，2022b）大幅减少了人力劳动，并使这些价值观更加明确：借助这项技术，只需将一组规范与价值观写成一份名为宪法的约束条件清单，就能训练模型遵循这些规范与价值观。此类技术还能大幅减少明确列出的广为认知的偏见示例，例如模型行为中的反黑人种族主义（Ganguli 等人，2023年）。[5]事实上，在某些情况下，在预训练期间让模型接触更多不良行为的示例，反而能让它们在部署时更容易避免这类行为，从而颠覆了训练数据与模型行为之间看似直观的联系（Korbak 等人，2023年；另见Chung 等人，2022年中的附录 C）。

这些技术干预措施，尤其是宪法式人工智能，易受外部影响和监管。人们很容易设想，第三方标准机构收集有关人工智能系统中哪些行为可接受的意见，并将这些意见提炼为模型开发者应当或必须采纳的宪法。

正如第4节中所述，不过，这些技术仍可能以微妙且出人意料的方式失效，而且随着模型规模扩大，这些技术的变化趋势十分复杂。当然，随着大规模人工智能系统的开发和部署，还会出现许多其他伦理问题，包括环境影响、可及性、滥用、隐私、安全性以及权力集中等问题（Amodei 等，2016年；Bender 等，2021年；Bommasani 等，2021年；Birhane 等，2022年；Weidinger 等，2022年，等）。

8. 与大语言模型的简短互动往往具有误导性

尽管许多已部署的大语言模型在很大程度上能够遵循指令，但这种指令遵循行为并非模型固有的，而是借助高度不完善的工具（第4节）附加到模型上的。部分由于这一点，模型会以各自特有的方式对指令内容表现出敏感性。模型常常会在收到任务要求时无法完成任务，但只要稍微改写或重新组织请求，就能正确完成任务，因此催生了提示工程(Brown 等人，2020年；Reynolds 和 McDonell，2021年；Radford 等人，2021年；Dohan 等人，2022年；White 等人，2023年；Si 等人，2023年）。

这些偶发性失败表明，我们用于控制语言模型遵循指令的技术并不总是有效。然而，仅仅观察到某个大语言模型在某种情境下无法完成某项任务，并不能可靠地证明该大语言模型不具备完成该任务所需的技能或知识。通常，一旦找到恰当的方式来提示模型完成某项任务，就会发现该模型始终能在该任务的不同实例中表现良好。第2节提到的思维链推理策略就是一个尤其清晰的例子：仅仅提示模型“逐步思考”，就能让它在整类数学和推理问题上表现良好，而在没有这一提示时，它本来会在这些问题上失败（Kojima 等人，2022年）。同样，即使观察到某个大语言模型始终无法完成某项任务，也远远不足以证明其他大语言模型都无法完成该任务（Bowman，2022年）。

另一方面，观察到大语言模型在某个实例中成功完成一项任务，并不能有力地证明该大语言模型普遍具备完成这项任务的能力，尤其是当该示例是作为演示的一部分被挑选出来的（例如图 4中的独角兽）。大语言模型可以记住训练数据中的特定示例或任务解法，却不一定内化了推理过程，因此无法稳健地完成这些任务（例如参见 McCoy 等人，2019年；Magar 和 Schwartz，2022年）。

9. 讨论与局限

我在此补充的讨论以以上八项论断为基础，并对其加以背景说明；但其中一些内容更具推测性或主观性，也反映了未必广为认同的观点。

9.1. 我们应当预期，当前大语言模型的一些显著缺陷会得到大幅改善

幻觉，即大语言模型编造看似可信却不真实的说法，是当前系统的一项显著缺陷，也极大限制了其负责任的使用方式。不过，第 3 节讨论的一些近期研究结果表明，我们或许很快就能通过更好地利用模型已有的能力来缓解这一问题：大语言模型会在内部以相当高的准确率追踪哪些陈述属实，而且这种能力会随着规模扩大而提升（Burns 等人，2023年；Kadavath 等人，2022年）。

同样，正如第7节所述，近期方法可以大幅减少模型输出中的显性偏见和毒性，这在很大程度上是利用了这样一个事实：当被询问时，模型往往能够识别这些不良行为（Dinan 等人，2019年；Bai 等人，2022b；Ganguli 等人，2023年）。虽然这些缓解措施不太可能完全稳健，但随着相关技术不断改进，这些不良行为的普遍程度和突出程度很可能会逐渐降低。

不过需要明确的是，这些令人鼓舞的迹象并不意味着我们能够可靠地控制这些模型，第4 节中指出的问题依然存在。我们提出的部分解决方案

[5]然而，模型展现出种族主义语言或决策的明确实例，并不足以全面揭示这些系统的开发和使用如何与涉及种族等因素的偏见和权力结构相互作用（例如，参见Field 等人，2021年）。

可能会留下重要的失效模式。例如，直接尝试处理幻觉很可能会悄无声息地失败，使模型看起来比实际更值得信任，这是因为存在相关的“藏拙”问题：如果我们采用标准方法训练某个未来的大语言模型，使其如实作答，但该模型能够相当准确地预测人工数据工作者可能会核查哪些事实性主张，那么它很容易只在提出可能会被核查的主张时才如实作答。

9.2. 将会出现激励因素，促使人们部署能够灵活追求目标的大语言模型智能体

随着能力不断增强、世界内部模型日益准确且实用，大语言模型很可能能够承担越来越开放式的任务，包括制定并执行新颖计划，以优化其在世界中的行动结果（Chan 等人，2023年）。随着这些能力的发展，经济激励因素表明，我们应会看到大语言模型被部署到软件工程或商业战略等领域；这些领域的特点是结果可衡量、需要灵活规划，且标准与法规相对灵活。借助额外工具增强的大语言模型，还能将应用拓展至机器人学等具身领域（Sharma 等人，2022年；Driess 等人，2023年）。这类部署将使大语言模型越来越频繁地置身于由系统自身行动所造成的陌生情境中，进一步降低开发者预测和控制其行为的能力。这可能会增加简单错误的发生率，使这些系统在某些情境下无法胜任智能体的角色。但这也可能增加更危险错误的风险：系统仍能有效运作，却在策略性地追求错误目标（Krueger 等人，2020年；Ortega 等人，2021年；Chan 等人，2023年）。

9.3. 大语言模型开发者对开发内容的影响有限

由于大语言模型的许多重要能力具有涌现性且难以预测，因此开发者对未来大语言模型将具备哪些具体能力的影响相对有限。试图根据开发者的经济激励、价值观或个性来预测未来大语言模型的能力，很可能会失败。例如，GPT-4似乎具备许多创建者可能期待的技能，例如编程相关技能。然而，它最初似乎也展现出几项不受欢迎的技能，例如教普通人制备生物武器，而创建者不得不投入大量精力试图将这些技能移除（OpenAI，2023年b）。

除此之外，大语言模型的开发者在决定是否部署某个大语言模型时，对其具备哪些能力也不可避免地了解有限：目前没有已知的评估或分析程序能够排除诸如思维链推理这类令人惊异的情况——在 GPT-3 中，用户发现了一种方法，可以诱发某种重要的新行为，而开发者此前并未意识到它的存在。

9.4. 大语言模型可能带来迅速增加的各类风险

更广泛地说，当前的技术和商业环境为快速构建和部署能力越来越强的大语言模型提供了强大的激励因素。尽管如此，在部署新型大语言模型之前，我们识别其能够展现哪些能力的记录并不稳定。我们控制系统的技术还很薄弱，应用于能力很强的模型时，很可能会进一步失效。鉴于此，我们有理由预期，大语言模型的开发和部署将带来大幅增加且性质显著变化的滥用风险和模型不当行为。

虽然基于大语言模型的系统有许多可能确实很有价值的积极应用，但若要事先评估其部署所涉及的社会成本效益权衡，可能仍然十分困难，甚至根本不可能；至少在模型评估、可解释性和控制等棘手技术问题取得重大进展之前是如此。某些难以评估的风险，例如涉及非常规武器或战略性权力寻求行为的风险，如果直到系统部署后才被发现，可能就无法得到充分缓解。尤其是，战略性权力寻求行为即使没有有意部署，也可能在模型开发期间造成严重风险。这表明，未来这方面的工作可能需要采用日益严格的安全性、安保和监督标准。

9.5. 大语言模型的负面结果可能难以解读，但能指出确实薄弱的领域

大量可靠的科学研究结果表明，即使在善意尝试引导其表现良好的情况下，近期的大语言模型在语言与常识推理任务上仍会出错，有时甚至连相对简单的任务也无法完成（Pandia 和 Ettinger，2021年；Schuster 和 Linzen，2022年）。有时，这些失败的具体表现会让人质疑其他相关评估的质量（Webson 和 Pavlick，2022年；Ullman，2023年）。鉴于第8节提到的原因，设计良好的测量指标所得到的正面结果，要比负面结果可靠得多。尽管如此，在一些领域中，包括处理否定这样简单的领域，[6] 大语言模型在语言处理或世界认知推理能力上，似乎存在系统性弱点。我们几乎没有依据预测这些局限是否会得到解决，或何时能够解决。

[6]例如，Huang 和 Wurgaft 提出的拒取式任务，见 McKenzie 等人(2022年)。

9.6. 关于大语言模型的科学研究与学术探讨尚不成熟

大语言模型给那些本应最有资格研究它们的领域带来了方法和范式上的挑战。自然语言处理（或语言技术）历来是这项工作的主要学科，但其工具旨在衡量并提升计算系统使用语言的能力。尽管大语言模型从根本上通过语言进行学习和交互，但有关其行为和能力的许多紧迫问题，主要并非语言使用方面的问题。研究人工智能政策和人工智能伦理的跨学科领域，已经建立了概念框架和规范性框架，用于思考各类人工智能系统的部署。然而，与近期大语言模型的实际情况相比，这些框架往往假设人工智能系统更精确地受其人类所有者和开发者的意图，或其训练数据的统计规律所支配（Chan 等人，2023年）。此外，许多关于大语言模型、被引用次数最多的研究论文，包括不少提出新方法或新理论的论文，都未发表于同行评审出版平台。近期限制大语言模型可及性、并将大语言模型训练的细节视为专有信息的趋势，也阻碍了科学研究。

这意味着，关于大语言模型的新颖而令人惊讶的论断，往往源自杂乱无章、存在谬误且超出既定学科实践的科学研究。然而，当既有的传统共识应用于大语言模型时，也往往根基不稳。所有这些都说明，对于本文讨论的问题，我们尤其应保持不确定，并以能够抵御错误假设的方式，就大语言模型作出重要决策。

结论

最后，我不打算重述上述论点，而想指出本文基本未涉及的三个有时备受关注的问题：

- 关于是否将大语言模型描述为理解语言，以及是否用知道或尝试等能动性相关词语来描述其行为，仍存在诸多争论；这些问题与本文讨论的问题大体无关（Bender& Koller，2020年；Michael，2020年；Potts，2020年）。无论这些系统在根本上是否具有人类特质——也就是这些词语所暗示的含义——我们都可以评估它们是有效还是无效、可靠还是不可靠、可解释还是不可解释，以及进步迅速还是缓慢。

- 同样，关于**意识**、大语言模型的感知能力、权利和道德受体地位的问题（例如，参见Schwitzgebel& Garza，2020年；Shevlin，2021年；Chalmers，2023年），也值得与上述问题区分开来。尽管这些问题可能会影响关于如何构建和使用人工智能系统的重要决策，但评估此处提出的大多数或全部问题，应当无需对这些问题表明立场。

- 最后，关于大语言模型的**价值判断**不在本文讨论范围之内。我们正在见证的大语言模型快速发展究竟是好事，以及我们每个人应该如何应对，这些更广泛的问题涉及的考量远比本文所依据的技术文献所能触及的更为深刻、多元。

致谢

本文得益于在罗格斯大学 Critical AI 组织的 AI FUTURES 专题讨论会上的交流，以及与许多其他研究人员的讨论，包括 Ellie Pavlick、Jackson Petty、Owain Evans、Adam Jermyn、Eric Drexler、Ben Garfinkel、RichardNgo、Jason Wei、Helen Toner、Jeffrey Ladish、Leo Gao、Alex Lyzhov、Julian Michael、Adam Bales、Rick Korzekwa、Ben Mann、Alex Lawsen、Alex Tamkin、Anton Korinek 和 David Dohan。本文中我仅使用大语言模型来探索一些细微的措辞和框架选择。当然，所有错误和遗漏都由我本人负责。

本研究获得了 Eric and Wendy Schmidt（根据 Schmidt Futures 项目的建议提供）和 Open Philanthropy 的资金支持，并获得了 FAR.ai 的 Pablo Moreno 提供的实物编辑支持。本材料基于美国国家科学基金会资助项目（编号 1922658 和 2046556）所支持的研究。本文所表达的任何观点、研究结果、结论或建议均属于作者个人，不一定反映美国国家科学基金会的观点。

参考文献

Abdou, M., Kulmizev, A., Hershcovich, D., Frank, S., Pavlick, E., and Søgaard, A. 语言模型能否在没有接地的情况下编码感知结构？以颜色为例进行个案研究。载于 第 *25* 届计算自然语言学习会议论文集，第 109–132 页，线上，2021年11月。计算语言学协会。数字对象标识符：10.18653/v1/2021.conll-1.9。网址：https://aclanthology.org/2021.conll-1.9。

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M. 等。Flamingo：用于少样本学习的视觉语言模型。神经信息处理系统进展，35：23716–23736，2022年。

Amodei, D.、Olah, C.、Steinhardt, J.、Christiano, P.、Schul- man, J. 和 Man´e, D. 人工智能安全中的具体问题。*arXiv* 预印本 *1606.06565*，2016年。

Andreas, J. 将语言模型视为智能体模型。载于 计算语言学协会论文集：*EMNLP 2022*，第 5769–5779 页，阿布扎比，阿联酋，2022年12月。计算语言学协会。网址 https://aclanthology.org/2022.findings-emnlp.423。

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., Das-Sarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., 等。通过人类反馈强化学习训练有帮助且无害的助手。*arXiv* 预印本 *2204.05862*，2022a。

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., 等。宪法式人工智能：来自人工智能反馈的无害性。*arXiv* 预印本 *2212.08073*，2022b。

Bartz, D. 随着 ChatGPT 人气飙升，美国立法者开始关注。路透社，2023年。网址 https://www.reuters.com/technology/chatgpts-popularity-explodes-us-lawmakers-take-an-interest-2023-02-13/。

Bender, E. M. 和 Koller, A. 迈向自然语言理解：数据时代的意义、形式与理解。载于计算语言学协会第*58*届年会论文集，第 5185–5198 页，2020年。

Bender, E. M.、Gebru, T.、McMillan-Major, A. 和 Shmitchell, S. 随机鹦鹉的危害：语言模型会不会过于庞大？ 载于*2021*年*ACM*公平、问责与透明度会议论文集，FAccT '21，第 610–623 页，美国纽约州纽约市，2021年。计算机协会。国际标准书号 9781450383097。数字对象标识符：10.1145/3442188.3445922。网址 https://doi.org/10.1145/3442188.3445922。

Bengio, Y.、Goodfellow, I. 和 Courville, A. 深度学习。马萨诸塞州剑桥：麻省理工学院出版社，2017年。国际标准书号 9780262035613。

Bengio, Y.、Russell, S.、Musk, E.、Wozniak, S. 等人。暂停大型人工智能实验。生命未来研究所公开信，2023年。网址 https://futureoflife.org/open-letter/pause-giant-ai-experiments/。

Birhane, A.、Kalluri, P.、Card, D.、Agnew, W.、Dotan, R. 和 Bao, M. 机器学习研究中编码的价值观。载于*2022*年*ACM*公平性、问责制与透明度会议，第173–184页，2022年。

Biswas, S. ChatGPT与医学写作的未来。《放射学》，第223312页，2023年。Bolukbasi, T.、Pearce, A.、Yuan, A.、Coenen, A.、Reif, E.、Vi´egas, F. 和 Wattenberg, M. BERT可解释性的一种错觉。*arXiv*预印本 *2104.07143*，2021年。Bommasani, R.、Hudson, D. A.、Adeli, E.、Altman, R.、Arora, S.、von Arx, S.、Bernstein, M. S.、Bohg, J.、Bosse-lut, A.、Brunskill, E. 等。论基础模型的机遇与风险。*arXiv*预印本 *2108.07258*，2021年。

Bowman, S.《少报风险的危险：报告自然语言处理系统失效情况时保持谨慎的理由》。载于第*60*届计算语言学协会年会论文集（第*1*卷：长篇论文），第7484–7499页，爱尔兰都柏林，2022年5月。计算语言学协会。数字对象标识符：10.18653/v1/2022 .acl-long.516。网址 https://aclanthology.org/2022.acl-long.516。

Branwen, G. 内心独白（人工智能），无日期。网址 https://gwern.net/doc/ai/nn/transformer/gpt/inner-monologue/index。

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., 等。语言模型是少样本学习者。神经信息处理系统进展，33：1877–1901，2020年。

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., 等。通用人工智能的火花：对 GPT-4 的早期实验。*arXiv*预印本 *2303.12712*，2023年。

Burns, C., Ye, H., Klein, D. 和 Steinhardt, J. 无需监督即可发现语言模型中的潜在知识。载于 第十一届国际学习表征会议，2023年。网址https://openreview.net/forum?id=ETKGuby0hcs。

Capoot, A. 微软宣布向 ChatGPT 制造商 OpenAI 投资数十亿美元。*CNBC*，2023年。网址 https://www.cnbc.com/2023/01/23/microsoft-announces-multibillion-dollar-investment-in-chatgpt-maker-openai.html。

Chalmers, D. J. 大语言模型可能具有意识吗？ *arXiv* 预印本 *2303.07103*，2023年。

Chan, A. GPT-3 与 InstructGPT：人工智能伦理与行业中的技术反乌托邦主义、乌托邦主义和"情境化"视角。*AI and Ethics*，第 1–12 页，2022年。

Chan, A.、Salganik, R.、Markelius, A.、Pang, C.、Rajkumar, N.、Krasheninnikov, D.、Langosco, L.、He, Z.、Duan, Y.、Carroll, M. 等。日益具备自主性的算法系统所造成的危害。*arXiv* 预印本 *2302.10329*，2023年。

Chan, L., Garriga-Alonso, A., Goldowsky-Dill, N., Green- blatt, R., Nitishinskaya, J., Radhakrishnan, A., Shlegeris, B., and Thomas, N. 因果擦除：一种严格检验可解释性假设的方法。*AlignmentForum*，2022年。网址 https://www.alignmentforum.org/posts/JvZhhzycHu2Yd57RN/causal-scrubbing-a-method-for-rigorously-testing。

Choi, J. H., Hickman, K. E., Monahan, A., and Schwarcz, D. ChatGPT 上法学院。明尼苏达法律研究论文，23(03)，2023年。数字对象标识符：http://dx.doi.org/10. 2139/ssrn.4335905。

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., 等。PaLM：通过路径扩展语言建模。*arXiv* 预印本 *2204.02311*，2022年。

Christiano, P. 引出潜在知识。*Medium*，2022年。网址 https://ai-alignment.com/eliciting-latent-knowledge-f977478608fc。

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., 等。指令微调语言模型的规模扩展。*arXiv* 预印本 *2210.11416*，2022年。Collins, K. M., Wong, C., Feng, J., Wei, M., 和 Tenenbaum, J. B. 结构化、灵活且稳健：对大语言模型进行基准测试并加以改进，使其在分布外推理任务中表现得更像人类。载于*2022*年认知科学（*CogSci*）会议，2022年。Devlin, J., Chang, M.-W., Lee, K., 和 Toutanova, K. BERT：用于语言理解的深度双向变换器预训练。载于计算语言学协会北美分会*2019*年会议论文集：人类语言技术，第*1*卷（长篇与短篇论文），第4171–4186页，明尼阿波利斯，明尼苏达州，2019年6月。计算语言学协会。数字对象标识符：10.18653/v1/N19-1423。网址 https://aclanthology.org/N19-1423。Di Langosco, L. L., Koch, J., Sharkey, L. D., Pfau, J., 和 Krueger, D. 深度强化学习中的目标误泛化。载于国际机器学习会议，第12004–12019页。PMLR，2022年。Dinan, E., Humeau, S., Chintagunta, B., 和 Weston, J. 为对话安全而构建、破坏并修复：通过对抗性人类攻击实现稳健性。载于论文集计算语言学协会*2019*年自然语言处理实证方法会议与第*9*届自然语言处理国际联合会议（*EMNLP-IJCNLP*）论文集，第4537–4546页，中国香港，2019年11月。计算语言学协会。数字对象标识符：10.18653/v1/D19-1461。网址 https://aclanthology.org/D19-1461。

Dohan, D., Xu, W., Lewkowycz, A., Austin, J., Bieber, D., Lopes, R. G., Wu, Y., Michalewski, H., Saurous, R. A., Sohl-Dickstein, J., 等。语言模型级联。发表于*Beyond Bayes*研讨会（*ICML 2022*），2022年。

Driess, D., Xia, F., Sajjadi, M. S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., 等。PaLM-E：一种具身多模态语言模型。*arXiv* 预印本 *2303.03378*，2023年。

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., 等。欠规格化给现代机器学习的可信度带来了挑战。《机器学习研究杂志》，2020年。

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., 等。一种用于 Transformer 电路的数学框架。*TransformerCircuits Thread*，2021年。网址 https://transformer-circuits.pub/2021/framework/index.html。Feng, S., Wallace, E., Grissom II, A., Iyyer, M., Rodriguez, P. 和 Boyd-Graber, J. 神经模型的病理特征使解释变得困难。载于*2018*年自然语言处理实证方法会议论文集，第 3719–3728 页，比利时布鲁塞尔，2018年10月至11月。计算语言学协会。数字对象标识符：10.18653/v1/D18- 1407。网址 https://aclanthology.org/D18-1407。Field, A., Blodgett, S. L., Waseem, Z. 和 Tsvetkov, Y. 自然语言处理领域关于种族、种族主义与反种族主义的综述。载于第*59*届计算语言学协会年会暨第*11*届国际自然语言处理联合会议论文集（第*1*卷：长篇论文），第 1905–1925 页，线上，2021年8月。计算语言学协会。数字对象标识符：10.18653/v1/2021.acl-long.149。网址 https://aclanthology.org/2021.acl-long.149。

Ganguli, D., Hernandez, D., Lovitt, L., Askell, A., Bai, Y., Chen, A., Conerly, T., Dassarma, N., Drain, D., Elhage, N., 等. 大型生成模型的可预测性与惊异。载于 *2022* 年 *ACM* 公平、问责与透明度会议，第 1747–1764 页，2022a。

Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., 等. 对语言模型进行红队测试以减少危害：方法、规模扩展行为与经验教训。*arXiv* 预印本 *2209.07858*，2022b。Ganguli, D., Askell, A., Schiefer, N., Liao, T., Lukoˇsi¯ut˙e, K., Chen, A., Goldie, A., Mirhoseini, A., Olsson, C., Hernandez, D., 等. 大语言模型的道德自我纠正能力。*arXiv* 预印本 *2302.07459*，2023年。

Gilkerson, J.、Richards, J. A.、Warren, S. F.、Montgomery, J. K.、Greenwood, C. R.、Kimbrough Oller, D.、Hansen, J. H. 和 Paul, T. D. 使用全天录音和自动化分析绘制早期语言环境图谱。《美国言语-语言病理学杂志》，26(2)：248–265，2017年。

Hart, B. 和 Risley, T. R. 对学习语言儿童的美国育儿方式：在自然家庭环境中观察到的家庭与儿童互动差异持续存在。发展心理学，28(6)：1096，1992年。

He, P.、Liu, X.、Gao, J. 和 Chen, W. DeBERTa：利用解耦注意力进行增强解码的 BERT。在 国际学习表征会议，2021年。网址 https://openreview.net/forum?id=XPZIaotutsD。

Hendrycks, D., Liu, X., Wallace, E., Dziedzic, A., Krishnan, R. 和 Song, D. 预训练变换器提升分布外鲁棒性。载于 第*58*届计算语言学协会年会论文集，第2744–2751页，线上，2020年7月。计算语言学协会。数字对象标识符：10.18653/v1/2020.acl-main.244。网址 https://aclanthology.org/2020.acl-main.244。

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Vinyals, O., Rae, J. W. 和 Sifre, L. 大语言模型训练的计算最优性实证分析。载于 Oh, A. H.、Agarwal, A.、Belgrave, D. 和 Cho, K.（编），神经信息处理系统进展，2022年。网址 https://openreview.net/forum?id=iBBcRUlOAPR。

Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., 和 Garrabrant, S. 先进机器学习系统中已学习的优化所带来的风险。*arXiv* 预印本 *1906.01820*，2019年。J, P. 和 C, D. ChatGPT 与大语言模型：风险何在？国家网络安全中心，2023年。网址 https://www.ncsc.gov.uk/blog-post/chatgpt-and-large-language-models-whats-the-risk。Jain, S. 和 Wallace, B. C. 注意力并非解释。载于*2019*年北美计算语言学协会会议论文集：人类语言技术，第*1*卷（长篇与短篇论文），第3543–3556页，明尼阿波利斯，明尼苏达州，2019年6月。计算语言学协会。数字对象标识符：10.18653/v1/N19-1357。网址 https://aclanthology.org/N19-1357。Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Dodds, Z. H., DasSarma, N., Tran-Johnson, E. 等。语言模型（大多）知道自己知道什么。*arXiv* 预印本 *2207.05221*，2022年。

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J. 和 Amodei, D. 神经语言模型的缩放定律。*arXiv*预印本 *2001.08361*，2020年。

Kenton, Z., Everitt, T., Weidinger, L., Gabriel, I., Mikulik, V. 和 Irving, G. 语言智能体的对齐。*arXiv*预印本 *2103.14659*，2021年。

Klein, E. 这改变了一切。《纽约时报》，2023年。网址 https://www.nytimes.com/2023/03/12/opinion/chatbots-artificial-intelligence-future-weirdness.html。

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y. 和 Iwasawa, Y. 大语言模型是零样本推理器。发表于 *ICML 2022*知识检索与语言模型研讨会，2022年。网址 https://openreview.net/forum?id=6p3AuaHAFiN。

Korbak, T., Shi, K., Chen, A., Bhalerao, R., Buckley, C. L., Phang, J., Bowman, S. R., and Perez, E. 使用人类偏好对语言模型进行预训练。*arXiv* 预印本 *2302.08582*，2023年。

Krueger, D.、Maharaj, T. 和 Leike, J. 自动诱发的分布偏移中的隐藏激励。*arXiv* 预印本 *2009.09153*，2020年。Levesque, H.、Davis, E. 和 Morgenstern, L. Winograd模式挑战。载于第十三届知识表示与推理原则国际会议，2012年。Li, B. Z.、Nye, M. 和 Andreas, J. 神经语言模型中意义的隐式表征。载于第*59*届计算语言学协会年会暨第*11*届自然语言处理国际联合会议论文集（第*1*卷：长篇论文），第1813–1827页，线上，

2021年8月。计算语言学协会。数字对象标识符：10.18653/v1/2021.acl-long.143。网址 https://aclanthology.org/2021.acl-long.143。

Li, K., Hopkins, A.K., Bau, D., Vi´egas, F., Pfister, H., and Wattenberg. 涌现的世界表征：探索在合成任务上训练的序列模型。载于第十一届国际学习表征会议，2023年。网址https://openreview.net/forum?id=DeG07TcZvT。

Li´etard, B., Abdou, M. 和 Søgaard, A. 语言模型知道通往罗马的路吗？载于 第四届 *BlackboxNLP* 研讨会论文集：分析和解读用于自然语言处理的神经网络，第 510–517 页，多米尼加共和国蓬塔卡纳，2021年11月。计算语言学协会。数字对象标识符：10.18653/v1/2021.blackboxnlp-1.40。网址 https://aclanthology.org/2021.blackboxnlp-1.40。

Lieu, T. 我是一名会编程的国会议员。人工智能让我感到害怕。《纽约时报》，2023年。网址 https://www.nytimes.com/2023/01/23/opinion/ted-lieu-ai-chatgpt-congress.html。

Lipton, Z. C. 模型可解释性的神话：在机器学习中，可解释性这一概念既重要又模糊。*Queue*，16(3):31–57，2018年。Lovering, C. 和 Pavlick, E. 神经网络概念的单元测试。《计算语言学协会会刊》，10:1193–1208，2022年。Lund, B. D. 和 Wang, T. 关于 ChatGPT 的讨论：人工智能和 GPT 可能如何影响学术界与图书馆？《图书馆高科技新闻》，2023年。数字对象标识符：https://doi.org/10.1108/LHTN-01-2023-0009。Magar, I. 和 Schwartz, R. 数据污染：从记忆到利用。载于 计算语言学协会第 *60* 届年会论文集（第二卷：短篇论文），第 157–165 页，爱尔兰都柏林，2022年5月。计算语言学协会。数字对象标识符：10.18653/v1/2022.acl-short.18。网址 https://aclanthology.org/2022.acl-short.18。McCoy, T.、Pavlick, E. 和 Linzen, T. 以错误的理由得出正确答案：诊断自然语言推理中的句法启发式。载于 计算语言学协会第 *57* 届年会论文集，第 3428–3448 页，意大利佛罗伦萨，2019年7月。计算语言学协会。数字对象标识符：10.18653/v1/P19-1334。网址 https://aclanthology.org/P19-1334。McKenzie, I.、Lyzhov, A.、Pieler, M.、Parrish, A.、Prabhu, A.、Mueller, A.、Kim, N.、Bowman, S. 和 Perez, E. 反向缩放奖：第二轮获奖者，2022年。网址 https://irmckenzie.co.uk/round2。

Mehdi, Y. 以人工智能驱动的全新微软必应和 Edge 重新定义搜索：它们是你的网络助手。微软官方博客，2023年。网址 https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/。

Menick, J., Trebacz, M., Mikulik, V., Aslanides, J., Song, F., Chadwick, M., Glaese, M., Young, S., Campbell-Gillingham, L., Irving, G. 等。教语言模型用经核实的引文支持答案。*arXiv* 预印本 *2203.11147*，2022年。

Michael, J. 剖析一只章鱼：理解形式与意义之争。博客文章，2020年。网址 https://julianmichael.org/blog/2020/07/23/to-dissect-an-octopus.html。

Michael, J., Holtzman, A., Parrish, A., Mueller, A., Wang, A., Chen, A., Madaan, D., Nangia, N., Pang, R. Y., Phang, J., 等。自然语言处理研究人员的看法是什么？自然语言处理社群元调查的结果。*arXiv* 预印本 *2208.12852*，2022年。Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., 等。WebGPT：借助浏览器并利用人类反馈进行问答。*arXiv* 预印本 *2112.09332*，2021年。Ngo, R. 从深度学习的角度看对齐问题。*arXiv* 预印本 *2209.00626*，2022年。Nye, M., Andreassen, A. J., Gur-Ari, G., Michalewski, H., Austin, J., Bieber, D., Dohan, D., Lewkowycz, A., Bosma, M., Luan, D., 等。展示你的工作：用于语言模型中间计算的草稿板。*arXiv* 预印本 *2112.00114*，2021年。Oliver, J.《约翰·奥利弗上周今夜秀》：2023年2月26日。网址 https://www.hbo.com/last-week-tonight-with-john-oliver/season-10/2-february-26-2022。

OpenAI。ChatGPT 插件，2023a。网址 https://openai.com/blog/chatgpt-plugins。OpenAI。GPT-4 技术报告。*arXiv* 预印本 *2303.08774*，2023b。网址 https://doi.org/10.48550/arXiv.2303.08774。Ortega, P. A., Kunesch, M., Del´etang, G., Genewein, T., Grau-Moya, J., Veness, J., Buchli, J., Degrave, J., Piot, B., Perolat, J., 等。动摇根基：序列模型在交互与控制中的妄想。*arXiv* 预印本 *2110.10819*，2021年。

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., 等。利用人类反馈训练语言模型遵循指令。神经信息处理系统进展，35:27730–27744，2022年。Pandia, L. 和 Ettinger, A. 梳理噪声：检验预训练语言模型的信息处理鲁棒性。载于自然语言处理实证方法会议，2021年。

Patel, R. 和 Pavlick, E. 将语言模型映射到具身概念空间。载于国际学习表征会议，2022年。

Perez, E., Ringer, S., Lukoˇsi¯ut ·
e,K., Nguyen, K.,Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., 等。通过模型撰写的评估发现语言模型的行为。*arXiv* 预印本 *2212.09251*，2022年。

Perrigo, B. 全新人工智能驱动的 Bing 正在威胁用户。这可不是闹着玩的。《时代》，2023年。网址 https://time.com/6256529/bing-openai-chatgpt-danger-alignment/。

Potts, C. 语言模型有可能实现语言理解吗？*Medium*，2020年。网址 https://chrisgpotts.medium.com/is-it-possible-for-language-models-to-achieve-language-understanding-81df45082ee2。

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. 等。通过生成式预训练提升语言理解。*OpenAI*博客，2018年。网址：https://openai.com/research/language-unsupervised。

Radford, A.、Wu, J.、Child, R.、Luan, D.、Amodei, D.、Sutskever, I. 等。语言模型是无监督的多任务学习者。*OpenAI*博客，2019年。网址：https://openai.com/research/better-language-models。

Radford, A.、Kim, J. W.、Hallacy, C.、Ramesh, A.、Goh, G.、Agarwal, S.、Sastry, G.、Askell, A.、Mishkin, P.、Clark, J. 等。从自然语言监督中学习可迁移视觉模型。载于国际机器学习会议，第 8748–8763 页。PMLR，2021年。

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. 和 Liu, P. J. 探索统一文本到文本转换器的迁移学习极限。《机器学习研究杂志》，21(1):5485–5551，2020年。

Reynolds, L. 和 McDonell, K. 面向大语言模型的提示编程：超越少样本范式。载于 *2021*年计算机系统人因会议（*CHI*）扩展摘要，第 1–7 页，2021年。Roose, K. 与 Bing 聊天机器人的一番对话让我深感不安。《纽约时报》，2023年。网址 https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html。

Saunders, W., Yeh, C., Wu, J., Bills, S., Ouyang, L., Ward, J. 和 Leike, J. 用于辅助人类评估者的自我批判模型。*arXiv* 预印本 *2206.05802*，2022年。

Schick, T., Dwivedi-Yu, J., Dess`ı, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N. 和 Scialom, T. Tool-former：语言模型可以自学如何使用工具。*arXiv* 预印本 *2302.04761*，2023 年。

Schuster, S. 和 Linzen, T. 当句子没有引入话语实体时，基于Transformer的模型有时仍会指代它。载于《*2022*年北美计算语言学协会分会年会论文集：人类语言技术》，第969–982页，美国西雅图，2022年7月。计算语言学协会。数字对象标识符：10.18653/v1/2022.naacl-main.71。网址 https://aclanthology.org/2022.naacl-main.71。

Schwitzgebel, E. 和 Garza, M. 设计具备权利、意识、自尊和自由的人工智能。载于《人工智能伦理学》。牛津大学出版社，2020年9月。国际标准书号 9780190905033。数字对象标识符：10.1093/oso/9780190905 033.003.0017。网址 https://doi.org/10.1093/oso/9780190905033.003.0017。

Sevilla, J.、Heim, L.、Ho, A.、Besiroglu, T.、Hobbhahn, M. 和 Villalobos, P. 机器学习三个时代的计算趋势。载于 *2022*年国际神经网络联合会议（*IJCNN*），第 1–8 页。IEEE，2022年。

Sharma, P.、Torralba, A. 和 Andreas, J. 基于潜在语言的技能诱导与规划。载于 第*60*届计算语言学协会年会论文集（第*1*卷：长篇论文），第 1713–1726 页，都柏林，爱尔兰，2022年5月。计算语言学协会。数字对象标识符：10.18653/v1/2022.acl-long.120。网址：https://aclanthology.org/2022.acl-long.120。

Shevlin, H. 我们如何知道机器人何时会成为道德受体？《剑桥医疗保健伦理季刊》，30（3）：459–471，2021年。数字对象标识符：10.1017/S0963180120001012。

Shlegeris, B.、Roger, F. 和 Chan, L. 语言模型在下一个词元预测方面似乎远胜于人类。*Alignment Forum*，2022年。网址https:

//www.alignmentforum.org/posts/htrZrxduciZ5QaCjw/language-models-seem-to-be-much-better-than-humans-at-next。

Si, C., Gan, Z., Yang, Z., Wang, S., Wang, J., Boyd-Graber, J. L., and Wang, L. 提示 GPT-3 变得可靠。载于第十一届国际学习表征会议，2023年。网址 https://openreview.net/forum?id=98p5x51L5af。

Silver,D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., 等。利用深度神经网络和树搜索掌握围棋。《自然》，529(7587):484–489，2016年。

Søgaard, A. 为章鱼的向量空间接地：从原始文本中理解词义。《心灵与机器》，第1–22页，2023年。Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., 等。超越模仿游戏：量化并外推语言模型的能力。*arXiv*预印本 *2206.04615*，2022年。Stein-Perlman, Z., Weinstein-Raun, B., 和 Grace, K. 2022年人工智能进展专家调查。*AI Impacts* 博客，2020年。网址 https://aiimpacts.org/2022-expert-survey-on-progress-in-ai/。Steinhardt, J. 关于基础模型中涌现行为风险的探讨。斯坦福 *CRFM* 博客文章，2021年。网址 https://crfm.stanford.edu/commentary/2021/10/18/steinhardt.html。Steinhardt, J. 人工智能预测：一年的回顾。*Bounded Regret*，2022年。网址 https://bounded-regret.ghost.io/ai-forecasting-one-year-in/。Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., 和 Christiano, P. F. 学习通过人类反馈进行摘要。神经信息处理系统进展，33: 3008–3021，2020年。Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozi`ere, B., Goyal, N., Hambro, E., Azhar, F., 等。LLaMA：开放且高效的基础语言模型。*arXiv*预印本 *2302.13971*，2023年。Turner, A. 和 Tadepalli, P. 可参数化重新指定目标的决策者往往会寻求权力。神经信息处理系统进展，35:31391–31401，2022年。Turner, A. M., Smith, L. R., Shah, R., Critch, A., 和 Tadepalli, P. 最优策略往往会寻求权力。载于 Beygelzimer, A., Dauphin, Y., Liang, P., 和 Vaughan, J. W.（编），神经信息处理系统进展，2021年。网址 https://openreview.net/forum?id=l7-DBWawSZH。Uesato, J.、Kushman, N.、Kumar, R.、Song, F.、Siegel, N.、Wang, L.、Creswell, A.、Irving, G. 和 Higgins, I. 基于过程和结果的反馈解决数学文字题。*arXiv*预印本*2211.14275*，2022年。Ullman, T. 大语言模型无法应对心智理论任务中的简单变体。*arXiv*预印本 *2302.08399*，2023年。Wang, B.、Min, S.、Deng, X.、Shen, J.、Wu, Y.、Zettlemoyer, L. 和 Sun, H. 探索理解思维链提示：关于关键因素的实证研究。*arXiv*预印本 *2212.10001*，2022年。Webson, A. 和 Pavlick, E. 基于提示的模型真的理解其提示的含义吗？载于计算语言学协会北美分会*2022*年会议论文集：人类语言技术，第2300–2344页，美国西雅图，2022年7月。计算语言学协会。数字对象标识符：10.18653/v1/2022.naacl-main.167。网址 https://aclanthology.org/2022.naacl-main.167。Wei, J.、Tay, Y.、Bommasani, R.、Raffel, C.、Zoph, B.、Borgeaud, S.、Yogatama, D.、Bosma, M.、Zhou, D.、Metzler, D.、Chi, E. H.、Hashimoto, T.、Vinyals, O.、Liang, P.、Dean, J. 和 Fedus, W. 大语言模型的涌现能力。机器学习研究汇刊，2022a。ISSN 2835-8856。网址 https://openreview.net/forum?id=yzkSU5zdwD。调查认证。Wei, J.、Wang, X.、Schuurmans, D.、Bosma, M.、brian ichter、Xia, F.、Chi, E. H.、Le, Q. V. 和 Zhou, D. 思维链提示能够在大语言模型中引发推理。载于 Oh, A. H.、Agarwal, A.、Belgrave, D. 和 Cho, K.（编），神经信息处理系统进展，2022b。网址 https://openreview.net/forum?id= VjQlMeSBJ。Weidinger, L.、Uesato, J.、Rauh, M.、Griffin, C.、Huang, P.-S.、Mellor, J.、Glaese, A.、Cheng, M.、Balle, B.、Kasirzadeh, A.、Biles, C.、Brown, S.、Kenton, Z.、Hawkins, W.、Stepleton, T.、Birhane, A.、Hendricks, L. A.、Rimell, L.、Isaac, W.、Haas, J.、Legassick, S.、Irving, G. 和 Gabriel, I. 语言模型带来的风险分类体系。载于*2022*年*ACM*公平性、问责制与透明度会议，FAccT '22，第214–229页，美国纽约州纽约市，2022年。计算机协会。国际标准书号9781450393522。数字对象标识符：10.1145/3531146.3533088。网址 https://doi.org/10.1145/3531146.3533088。

White, J.、Fu, Q.、Hays, S.、Sandborn, M.、Olea,C.、Gilbert, H.、Elnashar, A.、Spencer-Smith, J. 和 Schmidt, D. C. 使用ChatGPT增强提示工程的提示模式目录。*arXiv*预印本*2302.11382*，2023年。

Zhou, D., Sch¨arli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., Le, Q. V., and Chi, E. H. 由浅入深提示使大语言模型能够进行复杂推理。载于第十一届国际学习表征会议，2023年。网址https://openreview.net/forum?id=WZH7099tgfM。

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., andIrving, G. 基于人类偏好对语言模型进行微调。*arXiv* 预印本 *1909.08593*，2019年。