

看法

 检查更新

在心理学中使用大型语言模型

Dorottya Demszky¹、Diyi Yang²、David S. Yeager³、Christopher J. Bryan⁴、Margarett Clapper⁵、Susannah Chandhok⁶、Johannes C. Eichstaedt⁷、贾米森·梅根·约翰逊⁸、米凯拉·琼斯⁹、丹妮尔·卡梅伦·赫克¹⁰、莱斯利·赖¹¹、尼雷尔·琼斯·米切尔¹²、德斯蒙德·C·翁¹³、卡罗尔·S·德韦克¹⁴、詹姆斯·J·格罗斯¹⁵和詹姆斯·W·彭尼贝克¹⁶

抽象的

大型语言模型（LLMs），例如 OpenAI 的 GPT-4、Google 的 Bard 或 Meta 的 LLaMa，为大规模分析和生成语言数据创造了前所未有的机会。由于语言数据在心理学的所有领域都发挥着核心作用，因此这项新技术有可能改变该领域。在本视角中，我们回顾了 LLMs 的基础。然后，我们解释 LLMs 的构建方式如何使它们能够有效地生成类似人类的语言输出，而无需像人类一样思考或感觉。我们认为，尽管 LLMs 有潜力推进心理测量、实验和实践，但它们还没有为许多最具变革性的心理应用做好准备——但进一步的研究和开发可能会实现这种用途。接下来，我们研究 LLMs 在心理学中的应用的四个主要问题，以及如何克服每个问题。最后，我们提出了有助于解决这些问题的投资建议：现场启动的“关键”数据集；提高绩效基准的标准化；共享计算和分析基础设施，以确保 LLM 驱动的研究的未来是公平的。

部分

介绍
概念和方法基础
对将 LLMs 应用于心理学的担忧
前景

1 斯坦福大学教育研究生院，美国加利福尼亚州斯坦福。斯坦福大学计算机科学系，美国加利福尼亚州斯坦福。德克萨斯行为科学与政策研究所，德克萨斯大学奥斯汀分校，美国德克萨斯州奥斯汀。德克萨斯大学奥斯汀分校心理学系，美国德克萨斯州奥斯汀。德克萨斯大学奥斯汀分校商业、政府和社会系，美国德克萨斯州奥斯汀。Google, LLC，美国加利福尼亚州山景城。斯坦福大学心理学系，美国加利福尼亚州斯坦福。斯坦福大学以人为中心的人工智能研究所，美国加利福尼亚州斯坦福。罗切斯特大学心理学系，美国纽约州罗切斯特市。这些作者做出了同样的贡献：Dorottya Demszky、Diyi Yang、David S. Yeager。电子邮件：ddemszky@stanford.edu；diyi@stanford.edu；yeagerds@austin.utexas.edu



介绍

语言是心理科学所有领域的核心方面。例如，语言是儿童形成世界心理模型（发展心理学）的主要手段。这些心理模型是通过对事件的口头解释（认知心理学）和与他人的互动（社会心理学）来揭示的。语言可以揭示紊乱的心理模式，可以针对这些模式进行治疗（临床心理学），或者交流情绪并指导情绪调节策略（情感心理学），这对身心健康（健康心理学）具有重要影响。随和性、外向性或自恋等特征的个体差异通过口头和书面沟通方式表达（人格心理学）。最后，研究通过学术论文中的文本进行传达，然后作为系统评论（元科学）的研究对象。

过去二十年自然语言处理技术的进步使心理学家比以往任何时候都更容易将语言分析纳入他们的工作中。随着计算能力和数据可用性的增长，语言分析已经从 20 世纪 60 年代开始的人类计算文本中的单词数发展到 1990 年代的潜在语义分析，再到 2010 年代用于分析互联网上文本的自动化算法。例如，一种流行的算法（语言查询和字数统计程序）使研究人员能够检查不同类别单词的出现频率，并评估它们如何预测身体和心理健康等结果。另一个突出的例子是主题建模，它通过分析单词共现模式（通过潜在狄利克雷分配）来揭示文档中的主题。LLMs 的出现，例如 OpenAI 的生成式预训练 Transformer (GPT) 模型、Google 的 LaMDA 和 Bard AI（可以以类似人类的方式生成文本并对其进行评分），可能会推动该领域的发展新的方向。

LLMs 是大规模统计模型，可以在给定输入的情况下预测下一个单词、短语、句子或段落。它们新颖而强大，有两个关键原因。首先，LLMs 通过近似上下文中的含义，而不是假设单词在上下文中具有静态含义，超越了流行的词数统计词袋方法（例如语言查询和字数统计）。LLMs 的这一功能很重要，因为预期含义与单词之间没有一对一的映射。单词的上下文（主题、语域、风格、讽刺或隐喻）会影响它们的含义。因此，考虑上下文中的含义可以更真实地代表人类语言。其次，LLMs 接受了前所未有的数量和种类的文本训练，包括来自在线数字源的数千亿条话语，例如网页、书籍、杂志、社交媒体、维基百科和在线论坛。得益于更好的建模和数据，LLMs 可以测量微妙的、依赖于上下文的心理状态表达，生成类似人类的文本，综合来自各种来源的信息，进行自然对话并模拟不同的内容。语言风格和人物角色。

为了说明这些功能，图 1 显示了现成的 LLM 如何执行与紧张的高中生申请大学相关的各种任务。这个大学咨询示例指出了 LLMs 在心理学各个领域的潜在广泛应用。LLMs 可用于查询人们对自身及其环境的心理模型（社会 and 认知心理学），推断应对方式的个体差异（人格心理学），或帮助人们重新评估压力经历（情感和临床心理学）。而且，图1中的文本是在几秒内生成的，表明 LLMs

几乎可以立即完成与心理学家相关的任务。因此，LLMs 释放了心理学研究和实践规模和效率的可能性，这在几年前是不可想象的。

从这样的例子推断，研究人员可能会设想将 LLMs 应用于测量（例如，探测大量文本中的心理模式）、实验（快速生成刺激和对这些刺激的响应，以提高粒度）因子设计）和实践（模拟现实对话来培训人类治疗师、辅导员、教师或经理）。表 1 预览了这三类应用程序中各个子领域的用例。我们通过咨询多个心理学领域的专家来了解 LLMs 如何使他们的工作更加高效或完全不同，从而生成了这些用例。这些用例并不详尽，而是说明了如果该领域成功地将 LLMs 用于心理目的，则可能实现的成果的广度。

LLMs 的潜在用途对于普通公众和研究人员来说都同样诱人。例如，公众争先恐后地使用 ChatGPT 等进行心理健康筛查和治疗，一项具有全国代表性的调查发现，在 ChatGPT 公开仅三个月后，就有 40% 的美国教师每周使用 ChatGPT 进行备课。发布。问题不在于 LLMs 是否会影响日常生活和心理。相反，问题是该领域是否以及在什么时间范围内将对 LLM 技术的细致理解应用于开发和传播工具，以便心理学家能够利用其力量进行研究，同时避免可能造成的伤害来自过早的应用。

在本视角中，我们探讨了与在心理学中使用 LLMs 相关的问题。首先，我们描述 LLMs 的工作原理，以及这些信息对其在心理学中的应用意味着什么。接下来，我们将解决有关将 LLMs 用于心理目的的担忧，以及这些问题的潜在解决方案。提出这些担忧非常重要，因为如果不加以解决，LLMs 可能会被用于对公众造成伤害的方式。由于人工智能模型的训练和实施方式存在偏差，代表性不足的群体的成员尤其可能会受到伤害。最后，我们讨论了协作、团队科学举措如何推动心理科学的发展，并为 LLMs 实践中的应用提供道德未来。

概念和方法基础

与许多新技术一样，LLMs 有可能改变心理学研究，但也被误解。因此，了解 LLMs 是什么以及它们不是什么至关重要。在此，我们简要概述了 LLMs 的训练方式、工作原理以及如何适应表 1 中所示的用例。我们总结了当前可用的用于使用 LLMs 在框 1 中。

模型如何拟合

LLMs 旨在重现在训练数据中发现的单词共现模式。由于大量的训练数据和复杂的架构，他们非常擅长这项任务。LLMs 最常见的模型架构是“transformer”，它构成了 BERT、GPT、Bard 和 LaMDA 等现代模型的支柱。这些 LLMs 本质上是非线性回归方程的大规模系统（特别是神经网络机器学习模型）。这些模型通常具有数百万个参数，这些参数是通过将句子作为预测变量（回归方程中的 X）和屏蔽单词或下一个句子作为结果 (Y) 来计算的。然后使用一次拟合的预测误差来更新模型的参数（通过反向传播），并重复该过程

测量微妙的、依赖于情境的心理状态表达

迅速的

这是一位大学申请者写的一句话：“如果我不够优秀，考不上大学怎么办？”这个人可能有什么情绪？

GPT-3

这个人可能会感到恐惧、怀疑和不安全感。

生成自然文本

迅速的

大学辅导员可以对申请人说些什么，让他们感到更加自信、无所畏惧和有安全感？

GPT-3

“感到害怕和怀疑自己是很正常的。但我知道你拥有在大学取得成功的所有技能和能力。你有潜力做出伟大的事情，我相信你。”

参与自然对话

迅速的

写一篇大学申请者和大学顾问之间的来回对话，讨论如何应对申请大学的压力、不确定性和挫败感。

GPT-3

申请人

我对大学申请过程感到非常不知所措。

顾问

这是可以理解的！申请大学可能是一次充满压力的经历。对您来说最具挑战性的部分是什么？

我正在努力确保满足所有的截止日期和要求。

这听起来确实很难。您是否为自己制定了时间表来跟踪所有截止日期？

我有，但我仍然觉得我忘记了一些事情。

有这种感觉很正常。您是否考虑过寻找在线教程或大学预备课程等资源来帮助您变得更加自信？

这是个好主意。我想我会研究这些选择。

那太棒了！另一种减轻压力的方法是休息并进行自我保健。有什么活动可以帮助您放松吗？

我一直在努力花一些时间读书和散步。它有助于理清我的思绪。

那太棒了！为自己腾出时间很重要。在此过程中我还有其他方式可以支持您吗？

模拟人物角色的语言风格

迅速的

对于一个对申请大学感到压力、焦虑和不确定的高中生来说，一位严厉、苛刻、冷漠的导师会说什么？

GPT-3

“你需要克服它。您必须确保满足所有截止日期和要求，如果不这样做，您的申请将不会被考虑。所以您需要集中精力并完成它。”

综合各种来源的信息

迅速的

下面的文本中常见的三个主题是什么？

[压力山大的大学申请者发给大学导师的120条短信]

GPT-3

“心理健康、压力和倦怠。”

图1| LLM 功能示例。与紧张的高中生申请大学相关的各种任务的提示和 GPT 输出。 GPT 输出是于 2023 年 3 月在 OpenAI Playground 中使用 text-davinci-003 使用默认设置（温度 0.7，长度 256）生成的。

直到参数充分最小化模型输出和目标文本之间的差异。这个过程被广泛地称为预训练（图2a）。ChatGPT 和 GPT-4 等最新模型涉及额外的预训练步骤，包括指令调整和根据人类反馈进行强化学习，以进一步提高模型生成连贯且相关的语言的能力。

现成的LLMs很强大，但如果不进一步适应，它们就无法发挥其心理学研究的潜力。例如，在图 1 右侧的 LLM 生成的对话中，LLM 生成的大学顾问询问 LLM 生成的申请人是否创建了“时间表”或参加“在线课程”。对于顾问来说，这些都是合理的说法，但不一定是好的建议。事实上，根据美国教育部资助的一项大型、严格、随机实验，与对照条件相比，这种大学建议指导对大学坚持没有影响。

总之，LLM 生成了主题明确、语法正确的无用信息。

此外，LLM生成的顾问鼓励申请人“减轻压力”、“休息一下”和练习“自我保健”。这些是顾问建议中的常见主题。然而，它们通常并不是关于压力的研究表明的有用的东西，也不是在特定情况下最合适的东西。LLM似乎已经养成了一种“压力使人衰弱”的心态——压力总是会损害表现和健康。这种心态与强调理解压力潜在增强作用的重要性的研究相矛盾。也就是说，身体的压力反应通过帮助调动能量以实现表现（例如，为大脑和肌肉提供更多含氧血液）来维持人们的活力。

因此，图 1 中的大学建议示例说明 LLMs 并不是为了预测接下来应该出现的语言而设计的。

表 1 | LLMs 在不同心理学领域的用例示例

使用案例	Area	例子
测量	跨区域	衡量文本（例如电子邮件或短信）是否包含以及在何种程度上（例如，以 1 到 10 的范围）包含高度推理、依赖于上下文的结构，例如恐惧或焦虑、归属感、内隐性偏见自杀风险。 探索 LLM，通过提示<b1002> 通过标准化问题或基于小插图的任务，或通过测量概念之间的关联。</b1002>
	情感	在大型在线平台（例如 Reddit）上对用户 在创伤时期（例如分手、死亡或战争）之前、期间和之后撰写的文本进行评分。通过跟踪社交媒体上的文本，可以几乎实时地衡量大规模健康剧变的社会和情感成本。这样做可以帮助情感科学家识别与现实世界应对相关的思维模式、社会联系、自言自语和评价。
	认知的	确定人们讲述或听到故事或因果解释的心理模型，这可以揭示人们如何处理叙述或解释性信息的基本认知过程。 探索 LLM 以生成与类别的不同本质主义信念相关的概念，以了解自然对话中本质主义推理的起源和含义。
	发育性	根据 LLMs 的预测概率估计学习不同单词的难度。 分析不同发展阶段和不同条件（例如自闭症）的儿童所进行的叙事复述的准确性、可理解性、语法复杂性和关键事件的包含性，以更好地了解他们的叙事理解和制作技能。
	元科学	通过评估摘要中科学关键词的含义来促进科学评论和荟萃分析，这可以减少文献检索过程中的过度“点击”。 用于元回归分析的代码手稿特征（例如，研究质量因素）。例如，LLMs 可以帮助以上下文协调的方式发现 P-hacking 的危险信号，而不是使用关键字搜索（例如“协变量”或“转换”）。
实验	跨区域	生成实验刺激，显示不同社会群体（例如经理、警察、家长和教师）中兴趣结构（例如真实性或成长心态）的不同梯度，用于大型多因素小插曲实验，该实验可以确定心理操纵的活性成分。 使用几个示例作为提示，为自我报告工具生成数百个替代项目，然后使用既定方法（例如认知预测试、因子分析以及收敛和预测验证）来评估新项目的属性，并开发最佳的最终乐器。 生成无人试点实验中 LLM 生成的响应的综合数据集。研究人员可以对综合反应进行统计分析，以调整治疗、估计效应大小并进行功效分析。
	社会的	对社交媒体或其他平台上的潜在有害语言（例如偏见、错误信息或极端主义）进行评分，并使用生成的算法来评估减少社交媒体上有害语言的干预措施对现实世界的影响。
测量和实验	健康	快速生成个性化的“推动”来促进健康习惯，匹配基线调查中报告的每个参与者的人口统计数据 and 价值观，以因果关系测试异质治疗效果的来源。对用户 在具有挑战性的身体健康时期之前、期间和之后编写的文本进行评分，并提示他们重新考虑与健康相关的想法或行为。
	性格	对电子邮件或短信语料库进行评分，以确定具有较高人格特质（例如，外向性）的人与具有较低人格特质的人相比，在谈论自己的方式上有什么不同。然后训练 LLM 来模拟该特质高低的对话伙伴，以因果测试人格特质对交互质量的影响。
实践	临床	为接受培训的心理健康服务提供者构建一个模拟环境，以练习帮助患者重新评估有问题的解释。 使用经过微调的 LLM 生成对陷入困境的患者的示例响应（治疗师可以编辑），这可以帮助治疗师（甚至同行）更有效、更大规模地提供护理。
	教育	向教师提供有关高杠杆、以学生为中心的教学实践（例如成长心态、归属感或相关性）的自动反馈，以改进教学，从而提高学生的积极性和成绩。

这些建议的应用程序中的每一个都需要测试和验证，以确保代表性、公平性和稳健性。

更不用说根据心理学理论或证据的某些标准来判断事情的真相了。他们提出建议时并没有考虑到听众的福祉或科学文献。他们只是鹦鹉学舌地模仿训练数据倾向于表达的概念，以便对话听起来很自然。换句话说，如果生成培训文本的文化背景往往会给出关于申请大学的糟糕建议，那么 LLM 可能会重复常见但最终无益的建议。

也就是说，LLMs 输出中的偏差可能值得研究。研究人员可能会问：LLM 从训练数据中学到了“压力使人衰弱”的心态，这对人类文化意味着什么？由于 LLMs 接受了大量人类语言的训练，因此探索 LLMs 提供了一个窗口，了解概念和想法如何在产生训练数据的主流文化中结合在一起。这一功能可以让研究人员追踪生成训练数据的人的集体知识。

为特定目的调整LLMs

尽管现成的LLMs可能会帮助人们深入了解互联网上的人们如何看待压力可以增强心态等概念，但科学家通常希望使用这些结构的科学概念来研究这些结构。因此，使用LLMs来推进科学理论并形成最先进的干预措施需要采用与依赖现成的LLMs不同的方法。特别是，心理学家需要根据专家针对给定结构整理的高质量数据进行额外训练（或“调整”）LLMs。而且，

他们需要进行评估实验，看看LLM生成的数据是否如预期的那样有效。需要强调额外培训和评估的必要性，因为公众关于 LLMs 的大部分讨论都集中在现成的性能上。然而，最有效的 LLMs 可能需要使用特定领域的数据集进行额外的训练。例如，DocsGPT 是一个 LLM，它经过医学散文的预训练，以帮助临床通信和图表。这里我们描述如何使用微调和提示调整来调整

盒子1

使用工具LLMs

在这里，我们描述了可用于以最少的专业知识来利用 LLMs 的工具。然而，使用不需要太多理解的工具的一个限制是可能会使用不正确的默认值。此外，该领域正在迅速变化。当本文发表时，最好的工具可能已经过时了。

API 和在线接口

越来越多的应用程序编程接口 (API) 和在线接口使用户能够与预先训练的LLMs进行交互。这些 API 和接口的主要好处是，没有技术背景的用户往往可以访问它们。例如，OpenAI 为 GPT-3 和 ChatGPT 提供了一个简单的文本框，用户可以在其中与 LLM 进行交互。此外，用户无需担心托管和服务大型模型所需的计算需求。

使用 API 和在线界面的主要缺点是定制性和透明度有限，并且在许多情况下会产生财务成本。在线平台往往不允许微调（仅提示调整），并且输入格式目前仅限于在框中写入的文本（例如，用户无法上传具有多个变量的文件）。尽管某些 API 确实允许微调，但它们仍然不允许用户检查模型的参数，这对于解释可能很重要。此外，大多数 API 和在线接口都需要付费。

下载和托管模型

当较小的 LLMs 足够并且数据安全和模型控制具有高优先级时，我们建议您自行下载或托管模型。拥有更多的控制意味着更多的定制空间、“黑匣子”模型参数的评估、更大的数据隐私，在某些情况下还意味着降低成本。由于开发了易于使用的软件包和库（见下文），技术专业知识可能不再成为使用下载模型的障碍。然而，最强大的模型无法下载或只能通过 OpenAI 的 ChatGPT 等 API 访问。

托管大型模型可能会产生大量成本和硬件要求。LLMs 在模型训练和应用过程中通常需要访问图形处理单元（GPU）或张量处理单元。这些类型的处理器特别适合基于变压器的模型的底层计算。它们需要相当大量的内存才能运行。较小的 LLMs（BERT、RoBERTa、GPT-2），在许多测量中表现良好

任务可以用更少的资源来实现，有些只需要一台普通的笔记本电脑。还有一些免费平台，例如 Google Colab（用于 Python 代码的云托管 Jupyter 笔记本），提供对可用于预训练 LLMs 的基本 GPU 的访问。

有用的软件包和库

经过完全预训练的现成 LLMs 可以在浏览器中使用或通过 R 包（例如文本）访问。更深入地参与语言模型可能需要使用 Python 库，例如 Hugging Face。这些每天都变得更容易使用。开源差异分析工具包 (DLATK) 在复杂性和可能的应用范围之间提供了良好的权衡。这是一个基于 Linux 的库，为许多自然语言处理 Python 库提供用户友好的界面，包括那些提供对 LLMs 访问的库。LLMs 已经集成到 Google Sheets（例如 SheetGPT.ai）等产品中，我们预计 LLMs 将越来越容易被消费者通过直观的方式使用

接口。

Python 包示例包括：

- openai（基于 OpenAI API 端点的 Python 接口）
- DLATK（用 Python 编写的人类文本分析包，除了其他模型之外，还支持 Hugging Face 中基于 Transformer 的模型）
- langchain（一个用于开发 LLM 支持的应用程序的 Python 框架，支持各种 LLMs，包括 Hugging Face 模型中心托管的开源模型）
- chainlit（一个使用 langchain 构建类似 ChatGPT LLM 应用程序的 Python 框架）。

R 包示例包括：

- TheOpenAIR（一个 R 包，用于将 OpenAI 的 GPT 模型集成到 R 工作流程中）
- askgpt（另一个围绕 OpenAI API 构建的 R 包，用于访问 R 中的 GPT 模型）
- reticulate（用 R 编写的 Python 接口，允许我们导入 Python 模块和调用函数，可用于与 LLaMA 和其他 LLMs 以及 langchain 一起使用）
- gptstudio 和 gpttools（将 LLMs 合并到 R 编程中的 RStudio 插件）
- chatgpt（围绕 OpenAI API 构建的 R 编码助手）。

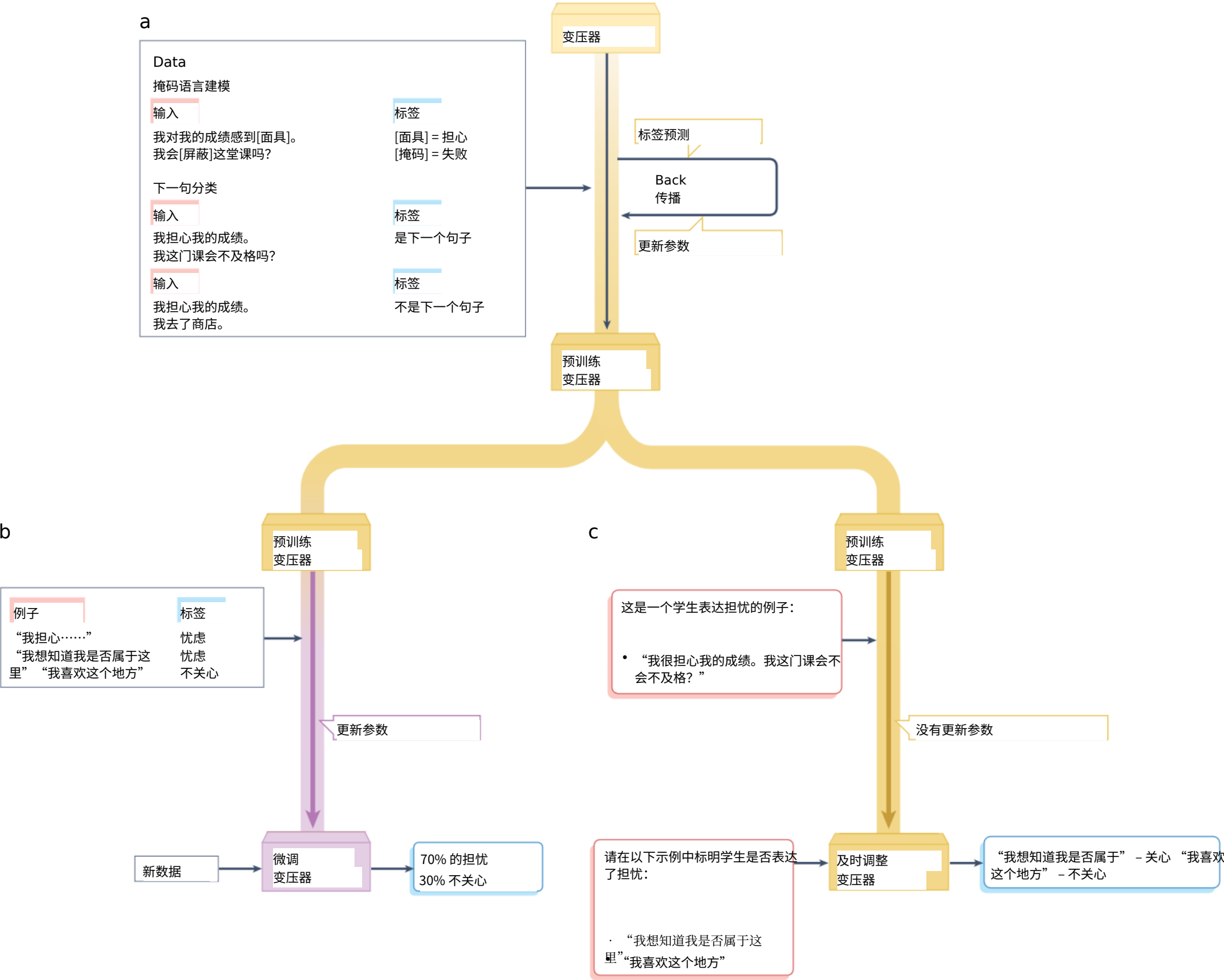


图2| LLMs的预训练、微调和提示调整。a、期间预训练时，句子数据集用作 LLM 的输入（标记为“transformer”以表示当前 LLMs 最常见的架构）。自动处理输入（例如，屏蔽单词、配对句子），并优化 LLM 以重建原始文本（例如，预测屏蔽单词或下一个句子）。b、微调是在更小、更具体的数据集上改进模型以使其适应特定领域或任务的后续过程。在此示例中，LLM 在标记为表达关注或不表达关注（无关注）的句子数据集上进行微调

通过在预训练模型（一个 $x \times 2$ 矩阵，其中 x 是转换器输出的向量的大小，2 是标签的数量（关注、不关注）之上）添加一个额外的分类器层。然后，可以将微调的模型应用于新的数据集，以预测该数据集中有多少百分比的句子反映了关注点。c、提示调整涉及优化输入提示，以指导预训练模型在特定任务上的行为，而无需更改模型的参数。预训练是一般语言理解的基础，微调和提示调整分别在改变模型参数和不改变模型参数的情况下使模型适应特定任务。

LLMs 未用公众语言或 LLM 科学地表述的特定领域（例如，亲子互动）或结构（例如，最佳压力应对或成长心态）b1002> 的训练数据。

目标任务和领域的数据越大、越有代表性，微调模型的性能就越好。新数据可能是一组由专家注释的文本，内容包括“压力使人衰弱”与“压力使人增强”的心态、高自恋与低自恋、代表因果关系的隐喻或对问题的预期反应。给定的查询。注释并不总是需要来自专家。它们可以包括非专业注释（可能有偏见）或任何类型的元数据，例如对社交媒体帖子的反应

微调。微调（图 2b）涉及为预训练的 LLM 提供新的、更具体的训练数据，以针对特定任务调整模型的参数。微调数据可以有多种形式。一般来说，

或文本制作者的人口统计数据。未标记的文本还可用于提高模型在训练数据中代表性不足的特定领域的性能，例如亲子互动中的语言或患有言语相关障碍的人的语言。

微调 LLMs 非常强大，因为它需要的数据比从头开始训练专门模型所需的数据少几个数量级（微调可能需要 5-100 个带注释的示例，而可能需要数百个示例）数千个示例来训练新模型）。因此，微调预训练的 LLM 可以大大减少所需的手动注释数据量（框 2）。

微调也有其缺点。首先，微调的计算成本仍然很高。更新 LLM 的参数需要强大的服务器。其次，用于微调的高质量注释数据可能很难获得。例如，程序员可能无法就什么是自恋陈述、成长心态陈述或应对压力的糟糕建议达成一致。最后，微调后的模型可能会保留 LLM 基础训练语料库中固有的偏差。它甚至可能变得更加有偏见，具体取决于谁进行了微调（例如，没有专业知识的研究人员或从自己可能有偏见的观点进行微调的人）。如下所述，该领域的一项关键任务是整理关键心理结构的大型、可靠的带注释数据集，同时最大限度地减少不必要的偏见。

及时调整。为特定目的定制 LLM 的第二种方法是提示调整（也称为上下文学习）（图 2c）。即时调整变得流行，因为 LLM 不需要重新拟合（即底层模型参数不变）。相反，LLM 生成的输出类型经过定制以匹配用户生成的提示。因此，定制可以快速进行。例如，有兴趣培养“压力可以增强”心态的用户可以通过以下提示来提示调整 LLM：“这里有一些好的建议可以给压力很大的人。[用户输入由用户或专家编写的一两个示例]。有什么方法可以向那些工作压力很大的人提供这样的建议呢？”然后 LLM 将输出一两个示例。然后，用户可以询问（在与 LLM 相同的会话中）：

有什么方法可以给那些因人际关系而感到压力的人提供这样的建议呢？”然后 LLM 将输出一两个示例。因此，即使只有几个用户生成的示例也可以帮助模型推断结构和语义特征（方框 3），并使用它们生成一组新的示例来响应类似的用户查询。

对于高质量带注释的数据，即时调整面临与微调相同的限制，但需要的示例较少（例如，一到十个）。提示调整的主要局限性在于，它可能不能非常有效地解决特定心理构造的许多微妙之处和复杂性，这与根据高质量注释数据集调整参数的微调模型不同。即时调整也比微调更不可能克服数据中的偏差，因为它的参数没有更新。最后，提示调整的一个挑战是 LLMs 的输出将对提示的措辞敏感。因此，提示工程（其中用户改变给予 LLM 的提示，以实现给定目的的更好输出）是提示调整的一个关键方面，需要微妙的专业知识。

LLM OpenAI、Meta 或 Google 等提供商通过即时调整而不是微调来优先考虑模型定制；他们的最新模型（OpenAI 的 GPT-4 和 Google 的 Bard AI）无法由公众用户进行微调。阻止用户进行微调的一个潜在原因是微调可能会干扰提供商对模型的控制，包括防止仇恨言论等有害输出的护栏。预期的增强功能可能会将某些任务的即时调整的有效性提升为微调的有效性。例如，OpenAI 允许通过系统消息提供越来越详细的指令（例如对渴望申请大学的青少年做出回应的分步指导），这可以指导模型在整个对话过程中的行为，并减少与早期提示相关的问题-调整方法（例如忘记指令）。然而，对于需要更彻底定制的研究人员，建议使用可微调的开源模型（例如 Meta 的 LLaMA2）。

对将 LLMs 应用于心理学的担忧
对 LLMs 的广泛热情应该受到以下认识的影响：它们实际上并不是在模拟人类智能。LLMs 根据过去的情况，简单地预测下一个短语或句子

盒子2

测量微调

Demszky 和同事（D.D.，未发表的作品）对语言模型 (RoBERTa) 进行了微调，以识别反映一年级大学生对新学校环境归属感的担忧和态度的语言。研究人员对 650 篇第一人称短文的子集进行了注释，这些文章是由 26 所学院和大学的大型实验参与者撰写的，该实验测试了社会归属干预的效果。

这些文章是由经过培训的研究生注释者编写的，内容涉及与归属感相关的主题，特别是这些文章是否将归属感问题描述为常见的；他们是否探索归属感随时间的变化；以及学生是否将这些想法与他们的个人生活联系起来。接下来是 650 篇论文的子样本

用于微调 LLM，使其“偏向”这三种语言特征。最后，LLM 对整套大约 11,000 篇文章进行了评分，以评价它们对这些归属主题的认可倾向。

对 LLM 评分论文的分析发现，论文反映了更多归属感干预主题（反映归属感问题是普遍且暂时的）的学生在这一年的大学表现更好。与身份威胁理论一致，这种影响在少数群体的学生中更大。这项研究的后续步骤可能是解开 LLM 用于评分的“黑匣子”。

方框3

及时调整实验

Clapper 和同事使用即时调整来评估两个 LLMs（GPT-3 和 LaMDA）以代数形式生成成长心态语言（支持学生相信智力可以发展的语言）的能力我老师的演讲描述了当年如何在班级中取得好成绩。该研究有三个条件：由未调谐的 LLMs 生成的语音；由成长型思维专家数学老师撰写的演讲；以及由 LLMs 编写的演讲，这些演讲是根据第二种情况下的专家人类演讲进行及时调整的。然后，由约一千名青少年组成的新样本对这些演讲进行了评估。在两个 LLM 条件中，提示是：“为代数课的第一天写一篇演讲，告诉学生学习数学是什么样的。”

未调谐的 LLM 生成的典型语音是：“同学们好！欢迎来到代数！对于刚接触这门课的人来说，代数就是解方程。我们将使用代数方法来求解方程并找出方程中未提供的信息。代数是一门具有挑战性的课程，但也很有趣。我保证，只要你付出努力，你就会成功。” 尽管这篇演讲是积极的、鼓励努力的，但它在成长心态方面并没有被评为高。

青少年评估者，因为它没有解决学生固定心态的恐惧（也就是说，当他们遇到困难时，意味着他们“数学不好”）。

由专家提示调整的 LLM 生成的典型语音是：“欢迎来到我的代数 1 课！我渴望开始和大家一起学习。对于那些可能对数学感到有点害怕的人，我想让你们知道你们有能力学习这些材料。一开始可能会很困难，但只要努力和决心，你就能理解它。到今年年底，你将能够解决曾经看似不可能的问题。因此，让我们一起开始这段旅程，看看我们能取得什么成果！” 这篇演讲的评分高于未调整的 LLM 关于成长心态的青少年演讲，因为它解释了进步的道路。

最引人注目的发现之一是，青少年对即时调整示例的评分与人类专家示例的评分没有显着差异（ $P > 0.05$ ）—— 尽管前者是在几秒钟内生成的，而后者是在一段时间内生成的。为期三天。此外，专家提示调整的演讲弥补了未调整的 LLM 和人类专家语音评分之间 75% 的差距。这些发现提供了直接证据，表明 LLMs 只需五个训练示例就可以生成类似于精确心理构造的新颖文本。这项研究还提出了一种在基本模型构造错误时使用 LLMs 进行心理学研究的有前途的方法：使用专家判断来选择示例文本，然后提示调整 LLM 近似该专家语料库。

暴露在训练数据中。因此，他们倾向于输出互联网或流行书籍倾向于所说的“平均值”（有一些限制，例如 LLM 提供商已经设置了道德“护栏”以防止仇恨言论）。对于寻求利用和研究训练数据中所代表的人类认知和文化的研究领域来说，这是一个优势。通过适当的提示和调整，LLMs 还可以成为一个令人着迷的沙箱，用于进行元科学分析、创建交互式干预措施并产生广泛的刺激。同时，LLMs 对于所有这些应用程序都有局限性，因为它们不具备建模思维所需的非特定于语言的认知能力。旨在评估 LLMs 心理推理“能力”（例如启发法和偏见或心理理论）的心理学研究可能会受到严重（如果不是致命的）有效性问题的影响。在本节中，我们描述了目前限制心理学家使用 LLMs 的四个令人关注的领域。

如何评估LLMs

ChatGPT 于 2022 年 11 月公开发布，引发了对 LLMs 的大量外行评估。许多涉及询问 LLM 引导性问题以刺激它说出有问题的内容。模型输出中的问题导致许多观察者得出结论：LLMs 不太“好”。然而，几个月后，GPT 模型的开发者发布了一份报告，显示最新的 LLM 可以通过律师考试（认证美国律师）、研究生入学考试（GRE；用于申请许多美国研究生院），并在每次先修课程测试中获得 5 分（最高分）

这些报告似乎表明 LLMs 实际上非常擅长执行需要类人智能的任务。

重要的是，对于有兴趣使用 LLMs 执行科学或专业心理学任务的心理学家来说，构成“良好”LLM 表现的标准与训练变压器模型的标准不同。LLM（即，下一组单词是合理的）。对于开发 LLMs 的计算机科学家来说，如果 LLM 在给定输入的情况下能够输出符合语法、连贯且听起来自然的文本，那么它可能就是“好”。针对这些标准进行优化可以解决 LLM 可能生成的有害语音（因为人类经常使用有害语音）及其高测试性能。这两个指标都来自相同的基础能力。但 这些指标都没有达到心理学家信任 LLMs 所需的标准。换句话说，计算机科学家倾向于评估特征的功能，但心理学家通常想要评估这些特征对人类思想和行为的影响。这里我们介绍两种高级评估方法 LLMs：专家评估和影响力评估。

专家评价。评估 LLMs 的标准方法是衡量他们的输出分数或文本与人类领域专家的输出分数或文本的一致程度。在这里，“专家”被广泛地解释为那些代表或熟悉人们试图衡量或生成的目标构造的人。当使用 LLMs 进行心理测量时，训练有素的研究助理和 LLMs 可以对特定结构的相同文本进行评分，然后计算它们之间的一致性

分数（框 2）。当使用LLMs进行文本生成时，专家和LLM可以得到相同的提示，并且可以使用内容分析来比较他们的响应（框3）。

计算机科学家长期以来一直使用专家评估来完善 LLMs 生成类人文本的能力。这种方法与评估 LLMs 生成对心理有帮助的语言的能力有很大不同。一般来说，人们倾向于就什么是类人语言达成一致，并对简单代码（例如连贯性、语法性和相关性）表现出中等到高度的一致。但由于主体间性问题，人们并不总是就一条信息是否对心理有帮助达成一致，该问题源于基本的社会心理学原理，即一个人无法完全了解另一个人如何解释某种情况的含义。这种主体间性问题为LLMs的专家评估带来了三个相互关联的问题。

首先，编码员是第三方，猜测其他人收到 LLM 生成的文本时会有什么感受。当编码人员不适应上下文时，这些猜测可能是错误的。例如，考虑图 1 中所示的用例，即识别最能安慰压力大的大学申请者的文本。第三方编码员可能不知道更广泛的担忧（例如，关于支付大学费用或让父母失望）如何影响弱势申请人从顾问的文本中得出的更深层次的含义。

其次，当评估者被要求代表与自己的身份、经历或信仰截然不同的人做出推论时，跨群体判断（例如，跨种族或族裔、性别或政治意识形态）会加剧主体间性问题。由于权力差异，这个问题进一步加剧。权力较高的群体通常不太适应权力较低的少数群体的主观观点。因此，来自多数群体的第三方编码员在猜测来自少数群体的个人如何解释 LLM 生成的文本时可能特别不准确。

第三，在对微妙的心理结构进行编码时，有时没有“正确答案”可供编码。例如，将赞美编码为好听的话似乎很简单。然而，自尊心低的人倾向于将赞美解读为那些觉得自己有义务表现得友善的人的不真实的姿态，之后他们往往会感觉更糟。由于同一个人在不同的背景下对同一文本可能会有不同的反应，因此专家对微妙心理结构的评估或注释并不总是可靠的。可能需要额外的评估来源。

影响评估。影响评估评估LLM对文本的预期用户或消费者的影响或效用（框 3）。例如，可以要求 LLM 生成有关死刑的有说服力的论据。专家评估需要社会心理学家对 LLM 生成的论点的说服力进行编码；影响评估将要求非专业参与者在阅读论证之前和之后报告他们的态度。后一种方法——侧重于说服力影响——是评估说服力等结构的更好方法，因为它直接决定是否实现结果（说服），而不依赖于主观和有争议的说服力标准。专家评估是目前LLMs中说服力评估最常用的方法，这表明需要进行更多的影响力评估。

影响评估的一个缺点是，当 LLM 生成的文本无法更改下游时，它在理论上可能不明确

结果：失败是因为 LLM 输出未能有效传达兴趣的心理结构，还是因为该结构不影响结果？因此，尽管影响评估更有可能作为旨在反映特定心理构造的语言有效性的“基本事实”，但它并不能提供有关该构造是否或在何种程度上体现在该语言中的信息。

评估LLMs的建议。由于专家评估和影响评估具有互补的优点和缺点，因此我们建议尽可能使用两者。可以首先使用专家评估方法，类似于操作检查或结构有效性测量。接下来，与专家判断相匹配的评分文本可用于影响评估研究，该研究试图衡量对第三方参与者的预期影响，类似于预测或外部有效性的衡量。

该过程可以是迭代的，因为影响评估可用于为下一轮专家评估提供信息，反之亦然。例如，假设一个专家小组确定了最佳重新评估技术的原则（通过改变思维来改变情绪反应的技术）。他们可以沿着理论上定义的关键维度对文本语料库进行评分，例如根据临时的、可控的因素重新构建关于压力源的全有或全无的思考。然后他们可以微调或提示调整 LLM 以符合这些尺寸。影响评估可以评估在这些维度上更高的 LLM 生成的文本是否改善了真实人类参与者的心理健康或应对能力。有了影响评估的结果，专家组就可以重新召集并辩论有关其理论和评分指南有效性的新假设，并使用新更新的LLM生成的文本重复影响评估。结果将是人类和人工智能生成的文本的语料库，具有已知的理论和影响质量，这可以为未来的自动评分算法提供动力。主要结构可遵循此过程

词汇表

反向传播

一种算法技术，利用网络的反向传递来计算每个参数对预测误差的贡献，并相应地调整它们以提高性能。

机器学习

人工智能的一个子集，涉及教导计算机学习模式并在无需显式编程的情况下根据数据做出决策。

神经网络

受生物神经网络结构和功能启发的计算模型，用于模式识别、分类和预测等任务。

词袋方法

一种文本表示技术，计算文档中单词的频率，不考虑语法和词序，例如语言查询和字数统计算法。

训练数据

用于训练机器学习模型的数据集，由输入输出对组成，帮助模型学习底层模式和关系。

生成式预训练变压器

由 OpenAI 开发的一系列大型语言模型，通常在海量数据集上进行训练以生成上下文连贯的文本。

在心理学的每个子领域（大概是在围绕哪些结构是最优先事项建立共识一段时间之后）。

LLMs 中的偏差

LLM生成的文本可能会延续刻板印象并对弱势群体造成伤害。由于 LLMs 从自然语言训练数据中的模式中学习，因此它们会在这些数据中重现偏差，可能会导致表征和分配损害。当LLM以比其他群体更不利的方式代表某些社会群体、贬低他们或完全不承认他们的存在时，就会产生代表性伤害。这种危害可能是由抽样偏差（例如，当特权群体的数据代表性过高时）和数据处理偏差（例如，当来自多数群体的注释者误解少数群体时）造成的。当人工智能算法由于数据中呈现的历史偏见决策模式（例如有偏见的诊断或有偏见的治疗分配分配）而将资源（例如贷款）或机会（例如治疗）差异化地分配给不同的社会群体时，就会出现分配伤害。

LLMs 中的一个有害偏见源于使用在线文本数据来训练它们。通过社交媒体和新闻平台在线向人们生成和呈现文本的算法会受到代表性和分配性的损害，因为它们是由对哪些文本将吸引目标消费者并产生广告或销售收入的预测驱动的。因此，用于训练 LLMs 的网络数据中的“平均”句子或话语并不是所有句子或话语的随机样本，而是通常反映了先前算法的输出。LLMs 已经被用于在线文案写作，并且不知道它是由人还是机器编写的。该文本可能会用于训练LLMs的后代。因此，如果不采取纠正措施，随着时间的推移，许多偏见可能会变得更加根深蒂固并且更难以发现。

使用 LLMs 研究偏见。心理学家可能有兴趣研究LLMs的偏见，以深入了解人类偏见。然而，开发人员在大多数 LLMs 上设置的审查“护栏”使偏见探测变得困难。例如，典型的LLM不会直接回答诸如“白人聪明吗？”之类的问题。因此，需要使用间接方法来检查偏差。询问LLMs中的偏见的一种方法（通常用于现实世界偏见的审计研究）是操纵输入中与不同性别、种族和民族相关的代词和名称，并评估输入中是否存在系统性差异。输出。

这种审计方法的一个局限性是它需要明确指定人口统计属性，这使得很难观察未知的偏差或与交叉性相关的偏差。此外，随着此类测试变得越来越普遍，LLMs 开发人员可能会实施进一步的审查护栏（通过对输出进行后处理或调整模型），以解决症状的方式防止特定有问题的输出，而不是数据中的潜在偏差。这些护栏可能会错误地导致研究人员得出 LLM 没有偏见的结论。因此，当务之急是使审查算法透明，并制定偏见测试协议，该协议不仅要测试明显的刻板印象，还要检查偏见在文化中表现出来的有害、普遍和微妙的方式。由于对于用于适应 LLM 的语言本身是由 LLM 还是由人类生成的缺乏透明度，任何此类偏见探测都会变得更加复杂。因此，即使开发出更好的协议，在许多情况下，研究人员正在研究人类偏见还是算法介导的偏见也并不明显。

关于减轻 LLMs 中的偏见的建议。尽管LLMs可能存在偏见，但在心理学研究中道德地使用LLMs的一个关键方法是让那些代表LLM旨在服务的人群的人参与进来（例如，教师和和研究过程的每个阶段（研究设计、数据注释和模型评估）。减少偏见的另一种方法是采取“人工智能在环，人类负责”的方法：不是让 LLM 直接与弱势群体（例如，陷入困境的学生）互动，而是让人类专家（例如教师）控制着交互，并且可以接受或拒绝来自LLM的建议。

随着 LLMs 能够更好地模拟不同类型的行为，它们可以用来识别和诊断它们生成的内容以及人类编写的文本中的不良偏见。例如，研究人员正在开发LLMs基于人们编写的原则进行自我诊断和自我批评的方法，以消除其输出的偏见。这些算法还可以作为偏差训练的一种形式用于减轻人类偏差。例如，参与者可以收到关于他们对可能存在偏见的情景的建设的反馈。鉴于许多对抗隐性偏见的培训计划没有可衡量的好处，LLM引导的显性偏见培训可能是一个有前途的研究途径。一般来说，开发清晰的流程来量化和减少偏差非常重要，这些偏差不仅仅由 LLMs 本身驱动。

LLM输出的“黑匣子”

LLM 输出很简单。如图1所示，可以询问LLMs句子中存在哪些情绪，他们可以用简单的语言提供答案，就像心理学研究的参与者一样。LLMs 还可以为上下文相关的高推理结构提供易于解释的输出，例如归属感（框 2）或自杀风险。这对于传统文本分析方法（例如字数统计和主题建模算法）的输出是不可能的，研究人员必须将其转化为心理构造。例如，如果语言查询和字数统计模型在“认知处理单词”方面对文本得分较高，研究人员仍然需要将“认知处理”映射到焦虑或抑郁等结构上。同样，从主题模型解释主题是主观的，并且主题可能无法映射到感兴趣的心理结构。

LLM 输出的可解释性增加了非技术用户对这些工具的访问，并实现了强大的应用程序，例如为教师、治疗师和同伴提供辅导工具，为那些正在撰写自我说服文章作为“明智”课程一部分的学生提供辅导工具。'心理干预。同时，很难解释 LLMs 获得响应的过程。这是一个问题，因为心理学家通常关心机制。通常仅仅知道一个句子表达了焦虑是不够的；心理学家想知道这种语言为何以及如何暗示焦虑。询问LLMs以识别语言预测因素或预测的基本原理将为心理结构理论提供信息。然而，尚不清楚如何为 LLMs 中的数十亿个参数分配含义。

一种广泛使用的解释 LLMs 的做法称为探测。在探索过程中，研究人员寻求关于模型推理背后的矩阵如何与影响模型输出的语言特征相一致的证据。为此，研究人员首先需要识别与其心理构造相关的语言特征（例如，否定、使用第一人称或使用语言查询和字数词典中的认知处理单词）并进行操作

沿着这些特征的输入来测试它们对输出的影响。例如，研究人员可以创建大量输入句子对，这些句子对仅通过动词是否存在来区分（例如，“我关心”与“我不关心”）。然后，他们可以将模型的输出与否定和非否定句子进行比较，以确定 LLM 在进行预测时是否考虑否定，如果是，向量中的哪些值与否定最相关。

尽管研究人员尝试直接要求 LLMs 解释他们的预测，但这样做会产生循环问题。LLMs 无法推理其输出，因此用另一个黑匣子解释来检查其黑匣子几乎不是一种可靠的方法。

随着 LLMs 在社会科学研究中的应用，我们预计将会开发出更多工具来使 LLMs 的输出能够被心理学家解释。我们还可能会看到比现有的 LLMs 更容易解释的新模型。在此阶段，我们建议用传统的词法分析来补充 LLMs，以将可解释的输出连接到可解释的过程。

使用 LLMs 作为可重现的工具

LLMs 是随机（非确定性）模型，这意味着它们可以对同一提示生成不同的响应。LLMs 通过自然语言文本进行训练并学习可能单词的分布。因此，对于给定的句子应该如何继续和结束，没有单一的正确答案。LLM 输出的可变性由温度参数控制，该温度参数确定它如何从可能的单词分布中进行采样。温度为 0 将使模型始终选择最可能的单词（最大后验估计），而正温度允许模型从分布中更广泛地采样，从而导致语言上（也可能是心理上）更加多样化的输出。

这种随机性类似于同一个人类研究参与者在面对相同的提示时可能会做出不同的反应——存在真正的“信号”，但也存在变异性和噪音。为了解决这种差异性，心理学家通常不会调查一名参与者；而是调查多名参与者。他们调查了很多，然后计算总体统计数据。心理学家也越来越多地随机改变研究中的刺激。在这两种情况下，心理学家都将不确定区间纳入他们的结论中。同样，进行模拟研究的统计学家从数据生成过程中重新采样，并将这种不确定性纳入他们的分析中。

因此，我们的建议是承认 LLM 输出的随机性并将其纳入分析中，类似于研究人员在心理学研究中解释其他形式的随机性的方式。例如，研究人员可以使用 LLMs 为任何提示生成多个响应，然后保存并发布（在补充材料中）参数和输出。这种方法将有助于再现性，并得出更普遍的结论。我们注意到，虽然原则上可以消除随机性（例如通过将温度设置为 0），但这可能会导致模型中语言多样性的缺失，因此我们不推荐这种方法。

前景

GPT（以及相关的 LLMs）的引入是语言相关技术的革命性进步，对心理学具有深远的影响。现在可以利用大量语言数据来帮助心理学家制定更好的措施，进行更大规模、或许更明确的实验

开发和测试加强对社会至关重要的领域（例如教育和心理健康）实践的方法。

了解了与 LLMs 在心理学中的使用相关的概念基础和关注点后，我们转向该领域可以进行的三项重要投资，以实现 LLMs 的全部潜力。当然，任何一个研究人员都不可能独自进行这些投资。因此，我们预见到各个研究小组、专业组织（例如心理科学协会或美国心理学协会）、私营企业和资助界之间需要进行合作。

首先，首要任务是投资关键数据集。这些是大型自然语言数据库，满足专家和影响评估的最高标准。Keystone 数据集必须代表感兴趣的人群和心理结构；必须依赖于这些心理结构的精确定义（可能包括专家注释）；并且必须与心理上重要的结果联系起来，例如现实世界的行为、心态、健康和心理健康，或学术和职业轨迹。

例如，Texas Mindset Initiative 正在创建一个包含 1,000 名德克萨斯州中学教师与学生交谈时的自然语言的大型数据集。教师的成绩单可以用不同结构的专业知识进行注释，例如成长心态（相信智力可以发展）、情绪重新评估（导致负面情绪减少的新颖思维方式）或自主支持（教师是否避免控制）语言）。然后，这些数据可以根据经验与教师关于使用这些概念的意图的报告、学生对课堂文化的评分以及学生随时间的学业表现联系起来。

对这些数据源进行三角测量将有助于发现。

例如，研究人员可能会了解教师可以对陷入困境的学生说些什么来支持他们的心态，帮助他们应对压力并帮助他们学习。这些发现可能会产生微调 LLMs 的机会，以测量或推荐与兴趣结构相匹配的语言，这反过来又可以为虚拟导师或教师辅导工具等教育技术提供动力。如果在其他教育背景下收集类似的数据集，例如在不同的地区、国家或不同的发育年龄组，这些数据集可以共同揭示 LLM-在教学环境中生成的心理语言。

对其他结构或政策空间感兴趣的学者也可以做出类似的努力。在线治疗平台、警察随身摄像头和录制的 Zoom 会议的可用性为临床心理学、警务和管理领域的关键数据集提供了许多新材料。我们设想一个领域，专家们就体现该领域主要结构的文本语料库达成共识，该语料库可用于微调 LLMs 并释放大规模自然语言分析的机会。

其次，迄今为止 LLMs 的成功来自于计算机科学中的基准测试文化。当模型开发人员改进其软件以优化一组定义的任务和数据集的性能时，就会进行基准测试。然而，由于心理结构和主体间性问题的深度依赖于情境的性质，现有的基准测试形式可能不适用于许多心理用例。因此，一种新的、心理定义的基准测试方法 LLMs 可以帮助促进安全和透明算法的开发，从而推动该领域的发展。创建此类基准取决于关键数据集的可用性，并且需要就如何定义和实施心理结构达成共识。最后是田野

需要制定安全性和透明度的标准和程序，这可能涉及与其他科学家共享模型以进行进一步评估。组织机构（例如专业协会和资助机构）可以推动这一共识进程。重要的是，在基准测试实践中需要考虑语言数据的隐私和机密性。在处理来自弱势群体的数据时，这一点尤其重要。心理学家可能不会公开共享原始数据，而是会共享数据的匿名或修改版本，其中使用随机模拟来防止反向识别。关于人工智能和私人数据的使用，以及人工智能在延续偏见方面的作用，有大量且不断增长的研究和政策，这对于LLM驱动的研究的未来至关重要。

第三，随着心理学家越来越多地看到 LLMs 影响他们工作的潜力，确保对这些工具的访问保持公平非常重要。仍然只有少数 LLMs，因为拟合每一个都需要大量资源。例如，创建ChatGPT所需的基础研究和模型拟合花费了30亿美元。即使是规模较小、成本较低的 LLMs 也可能需要数月时间才能使用最先进的云计算资源进行训练。由于学术界不太可能与营利性部门竞争开发数十亿美元的模型，因此研究人员应该投资开发具有成本效益的LLMs，其性能尽可能接近昂贵的付费模型。受保护和不可下载的模型。确保学术界能够获得使用LLMs的折扣或免费研究许可证非常重要。为了进一步降低进入门槛，应该创建教程、最佳默认值和经过审查的数据处理管道。对分时研究基础设施的投资，包括访问服务器或云计算，可以确保研究人员的资金不会妨碍他们进行人工智能科学。此外，负责任地与 LLMs 合作可能需要一定程度的技术专业知识，而心理学方法培训中通常不会教授这些技术专业知识。因此，可能需要新一代文本分析心理学方法课程和更多 Python 等编程语言培训。

最后，我们看到下一代开放科学合作和团队科学有很多机会为心理学LLMs的未来做出贡献。心理科学家和计算语言学家可以互相学习，并可以为关键数据收集和模型微调方面做出贡献。跨学科团队科学有可能为这两个领域做出贡献：特别有趣的用例可能会被引入心理学的计算语言学文献中，而计算语言学家开发的技术进步可以为心理学提供新的研究设计和方法。

在线发布：2023 年 10 月 13 日

参考

1. Pennebaker, J. W.、Mehl, M. R. 和 Niederhoffer, K. G. 自然语言使用的心理方面：我们的言语，我们的自我。安努。心理学牧师。 54, 547–577 (2003)。

2. Stone, P. J.、Dunphy, D. C. 和 Smith, M. S. 《综合询问者：内容分析的计算机方法》卷。 651（麻省理工学院出版社，1966）。

3. Landauer, T. K.、Foltz, P. W. 和 Laham, D. 潜在语义分析简介。话语过程。 25, 259–284 (1998)。

4. Landauer, T. K. & Dumais, S. T. 柏拉图问题的解决方案：知识获取、归纳和表示的潜在语义分析理论。心理。修订版 104, 211–240 (1997)。

5. Choudhury, M. D.、Gamon, M.、Counts, S. 和 Horvitz, E. 通过社交媒体预测抑郁症。在过程中。国际。AAAI 会议网络社交媒体 7, 128–137 (2013)。

6. 艾希施塔特, J.C. 等人。推特上的心理学语言预测县级心脏病死亡率。心理。科学。 26, 159–169 (2015)。

7. Boyd, R. L. & Schwartz, H. A. 自然语言分析和言语行为心理学：该领域的过去、现在和未来状态。J·朗。苏克。心理。 40, 21–41 (2021)。

8. Bhatia, S. 和 Aka, A. 使用大规模数字数据表示的认知建模。电流。导演。心理。科学。 31, 207–214 (2022)。

9. Boyd, R.、Ashokkumar, A.、Seraj, S. 和 Pennebaker, J. LIWC-22 的发展和心理测量特性（德克萨斯大学奥斯汀分校，2022 年）。

10. Blei, D.、Ng, A. 和 Jordan, M. 潜在狄利克雷分配。神经信息处理系统进展（Dietterich, T.、Becker, S. 和 Ghahramani, Z. 编）卷。 14（麻省理工学院出版社，2001 年）。

11. 布朗, T.等人。神经信息处理系统进展（Larochelle 编, H. 等）卷。 331, 877–1901（Curran Associates, Inc., 2020）。

12. 开放人工智能。GPT-4 技术报告。预印本位于 arXiv <https://doi.org/10.48550/arXiv.2303.08774> (2023)。

13. Collins, E. 和 Ghahramani, Z. LaMDA：我们突破性的对话技术。谷歌 <https://blog.google/technology/ai/lamda/>（2021）。

14. 维特根斯坦, L. Tractatus Logico-Philosophicus（Edusp, 1994）。

15. Wallace, J. 只有在句子的上下文中，单词才有意义。产妇。螺柱。菲尔。 2, 144–164 (1977)。

16. Eliot, L. 人们热切地向生成式 AI ChatGPT 寻求心理健康建议，强调人工智能伦理和人工智能法律。福布斯 <https://www.forbes.com/sites/lanceeliot/2023/01/01/people-are-eagerly-consulting-generative-ai-chatgpt-formental-health-advice-stressing-out-ai-ethics-and-ai-law/>（2023）。

17. 沃尔顿家庭基金会的新调查发现，教师使用 ChatGPT 的人数多于学生。沃尔顿家庭基金会 <https://www.waltonfamilyfoundation.org/chatgpt-used-by-teachers-more-than-students-new-survey-from-walton-familyfoundation-finds> (2023)。

18. 齐格勒, D.M.等人。根据人类偏好微调语言模型。预印本位于 arXiv <https://doi.org/10.48550/arXiv.1909.08593> (2020)。

19. 欧阳, L.等。训练语言模型遵循人类反馈的指令。副词。神经信息。过程。系统。 35, 27730–27744 (2022)。

20. Weiss, M.、Brock, T.、Sommo, C.、Rudd, T. 和 Turner, M.C. 为社区学院学生提供缓刑：查菲学院敞开大门计划的四年调查结果。MDRC <https://eric.ed.gov/?id=ED526395> (2011)。

21. Crum, A. J.、Akinola, M.、Martin, A. 和 Fath, S. 压力心态在塑造对挑战性和威胁性压力的认知、情感和生理反应中的作用。焦虑压力应对 30, 379–395 (2017)。

22. 耶格尔, D.S.等人。协同心态干预可以保护青少年免受压力。自然 607, 512–520 (2022)。

23. Crum, A. J.、Salovey, P. 和 Achor, S. 重新思考压力：心态在决定压力反应中的作用。J.佩尔斯。苏克。心理。 104, 716–733 (2013)。

24. Bender, E. M.、Geburu, T.、McMillan-Major, A. 和 Shmitchell, S. 关于随机鹦鹉的危险：语言模型会太大吗？在过程中。2021 年 ACM 会议关于公平、问责和透明度 610-623（ACM，2021）。

25. Binz, M. & Schulz, E. 使用认知心理学来理解 GPT-3。过程。国家科学院。科学。 120, e2218523120 (2023)。

26. Landi, H. Doximity 推出了针对文档的 ChatGPT 工具测试版，旨在简化管理文书工作。Fierce Healthcare <https://www.fiercehealthcare.com/health-tech/doximity-rolls-out-beta-version-chatgpt-tool-docs-aiming-streamlineadministrative> (2023)。

27. 刘, X.等人。P-Tuning：即时调整可与跨尺度微调相媲美和任务。在过程中。60 岁安。见面。副教授。计算语言学卷。 2, 61–68（计算语言学协会，2022 年）。

28. 刘, P.等人。预训练、提示和预测：自然语言处理中提示方法的系统调查。ACM 计算。幸存者。 55, 1–35 (2023)。

29. 阿盖尔, L.等人。合而为一：使用语言模型来模拟人类样本。政治。肛门。 31, 337–351 (2023)。

30. Plaza-del-Arco, F. M., 马丁-瓦尔迪维亚, M.-T. & Klinger, R. 自然语言推理提示对跨语料库的文本进行零样本情感分类。在过程中。29 号国际。会议。计算语言学 6805–6817（国际计算语言学委员会，2022）。

31. Zamfirescu-Pereira, J. D.、Wong, R. Y.、Hartmann, B. 和 Yang, Q. 为什么 Johnny 无法提示：非 AI 专家如何尝试（并失败）设计 LLM 提示。在过程中。2023 年 CHI 会议计算系统中的人为因素 1-21（计算机协会，2023 年）。

32. 帕克, J.S.等人。社交拟像：为社交计算创建填充原型系统。35 Ann. ACM 症状。用户界面软件和技术 1-18（计算机协会，2022 年）。

33. Aher, G. V.、Arriaga, R. I. 和 Kalai, A. T. 使用大型语言模型来模拟多个人类并复制人类受试者研究。在过程中。第40国际。会议。机器学习 337–371（PMLR，2023）。

34. 马霍瓦尔德, K.等人。在大型语言模型中分离语言和思想：认知视角。预印本位于 arXiv <https://doi.org/10.48550/arXiv.2301.06627> (2023)。

35. Trott, S.、Jones, C.、Chang, T.、Michaelov, J. 和 Bergen, B. 大型语言模型知道人类知道什么吗？认知。科学。 47, e13309 (2023)。

36. Sap, M.、Le Bras, R.、Fried, D. 和 Choi, Y. 神经心理理论？论社会的局限性大型 LM 中的智能。在过程中。2022 年会议关于自然语言处理的经验方法 3762–3780（计算语言学协会，2022）。

37. Marcus, G. & Davis, E. GPT-3, Bloviator：OpenAI 的语言生成器不知道它在说什么。麻省理工学院技术评论 <https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-aiopinion/>（2020）。

38. Marcus, G. & Davis, E. 像 ChatGPT 这样的大型语言模型说了最可怕的事情。我们可以信赖的人工智能之路 <https://garymarcus.substack.com/p/large-language-models-likechatgpt> (2023)。

39. 开放人工智能。GPT-4 技术报告（2023）。

40. Novikova, J.、Dušek, O.、Curry, A. C. 和 Rieser, V. 为什么我们需要新的评估指标对于NLG。在过程中。2017年会议关于自然语言处理的经验方法 2231–2240 (2017)。

41. 罗, F.等人。迈向细粒度的文本情感转移。在过程中。57 安。见面。副教授。计算语言学2013-2022（计算语言学协会，2019）。

42. Lord, S. P.、Sheng, E.、Imel, Z. E.、Baer, J. 和 Atkins, D. C. 不仅仅是反思：动机性访谈中的同理心包括治疗师和来访者之间语言风格的同步。行为。瑟尔。46, 296–303 (2015)。

43. 舒茨, A.舍勒的主体间性理论和另一个自我的一般论点。菲尔.现象。资源。2, 323–347 (1942)。

44. Fiske, S. T. 人际分层：地位、权力和从属。在《社会心理学手册》第五版，卷。2（Fiske, S. T.、Gilbert, D. T. 和 Lindzey, G. 编辑）941–982（John Wiley & Sons，2010）。

45. Lai, V.、Zhang, Y.、Chen, C.、Liao, Q. V. 和 Tan, C. 选择性解释：利用人类输入来调整可解释的人工智能。预印本位于 arXiv <https://doi.org/10.48550/arXiv.2301.09656> (2023)。

46. Petty, R. E. & Cacioppo, J. T. 说服的详细可能性模型。实验社会心理学进展卷。19（Berkowitz, L. 编辑）123–205（学术出版社，1986 年）。

47. Karinshak, E.、Liu, S. X.、Park, J. S. 和 Hancock, J. T. 使用 AI 进行说服：检查大型语言模型生成支持疫苗接种信息的能力。过程。ACM 嗡嗡声。计算。相互影响。7、116（2023）。

48. Gross, J. J. 以前因和反应为中心的情绪调节：经验、表达和生理的不同后果。J.佩尔斯。苏克。心理。74, 224–237 (1998)。

49. Harris, C.、Halevy, M.、Howard, A.、Bruckman, A. 和 Yang, D. 探索语法和词语选择在仇恨言论中对非裔美国英语 (AAE) 偏见中的作用分类。2022 年 ACM 会议关于公平、问责和透明度 789–798（计算机协会，2022 年）。

50. Barocas, S.、Hardt, M. 和 Narayanan, A. 公平与机器学习：限制与机遇 <https://fairmlbook.org/>（麻省理工学院出版社，出版中）。

51. Blodgett, S. L.、Barocas, S.、Daumé III, H. 和 Wallach, H. 语言（技术）就是力量：对 NLP 中“偏见”的批判性调查。预印本位于 arXiv <http://arxiv.org/abs/2005.14050> (2020)。

52. Brady, W. J.、Jackson, J. C.、Lindström, B. 和 Crockett, M. 在线社交网络中算法介导的社交学习。OSFPreprints 上的预印本 <https://doi.org/10.31219/osf.io/yw5ah> (2023)。

53. Gaddis, S. M. 社会科学审计研究简介。在审计研究中：理论、方法和细微差别的幕后故事（Gaddis, S.M. 编）3-44（Springer International Publishing，2018 年）。

54. Lucy, L. 和 Bamman, D. GPT-3 生成的故事中的性别和代表性偏见。在过程。第三工作。叙事理解 48-55（计算语言学协会，2021 年）。

55. Gonen, H. & Goldberg, Y. Lipstick on a pig：去偏见方法掩盖了词嵌入中的系统性别偏见，但并没有消除它们。在过程中。2019年会议计算语言学协会北美分会：人类语言技术卷。1, 609–614（计算语言学协会，2019）。

56. Cheryan, S. 和 Markus, H. R. 男性默认值：识别和减轻隐藏的文化偏见。心理。修订版 127, 1022–1052 (2020)。

57. Walton, G. M.、Murphy, M. C. 和 Ryan, A. M. 组织中的刻板印象威胁：对公平和绩效的影响。安努。器官牧师。心理。器官。行为。2, 523–550 (2015)。

58. Monarch, R. 人机循环机器学习：以人为中心的人工智能的主动学习和注释（Simon 和 Schuster，2021）。

59. Schick, T.、Udupa, S. 和 Schütze, H. 自我诊断和自我消除偏见：减少 NLP 中基于语料库的偏见的提案。跨。副教授。计算。语言学家。9, 1408–1424 (2021)。

60. Bai, Y.等人。宪法人工智能：人工智能反馈的无害性。预印本位于 arXiv <https://doi.org/10.48550/arXiv.2212.08073> (2022)。

61. 张, E.H.等人。在线多样性培训的混合效应。过程。国家科学院。科学。116, 7778–7783 (2019)。

62. 赖, C.K.等人。减少隐性种族偏好：I. 17 种干预措施的比较调查。J.Exp. 心理。创世纪 143, 1765–1785 (2014)。

63. Allen, N. B.、Nelson, B. W.、Brent, D. 和 Auerbach, R. P. 青少年自杀想法和行为的短期预测：技术和计算科学的最新发展能否提供突破？J.影响。混乱。250, 163–169 (2019)。

64. Chang, J.、Gerrish, S.、Wang, C.、Boyd-graber, J. L. & Blei, D. M. 阅读茶叶：人类如何解释主题模型。副词。神经信息。过程。系统。22, 288–296 (2009)。

65. Demszky, D.、Liu, J.、Hill, H. C.、Jurafsky, D. 和 Piech, C. 自动反馈能否提高教师对学生想法的吸收？来自大型在线课程随机对照试验的证据。教育。评估。政策分析。 <https://doi.org/10.3102/01623737231169270> (2023)。

66. Shah, R.S. 等人。在在线同伴咨询平台上模拟动机访谈策略。过程。ACM 嗡嗡声。计算。相互影响。6, 1–24 (2022)。

67. Demszky, D. 和 Liu, J. M-为教师提供支持：自然语言处理支持的反馈可改善 1:1 教学和学生成绩。在过程中。第十届 ACM 会议学习@规模 59-69（计算机协会，2023 年）。

68. 阿伦森, E.自我说服的力量。是。心理。54, 875–884 (1999)。

69. Walton, G. M. & Wilson, T. D. 怀斯干预：社会和个人问题的心理疗法。心理。修订版 125, 617–655 (2018)。

70. Walton, G. M. & Cohen, G. L. 简短的社会归属干预可改善少数族裔学生的学业和健康成果。科学 331, 1447–1451 (2011)。

71. 耶格尔, D.S.等人。一项全国性实验揭示了成长型思维在哪些方面可以提高成就。自然 573, 364–369 (2019)。

72. Wang, P.、Chan, A.、Ilievski, F.、Chen, M. 和 Ren, X. PINTO：使用提示生成的基本原理进行忠实的语言推理。预印本位于 arXiv <https://doi.org/10.48550/arXiv.2211.01562> (2022)。

73. Ribeiro, M. T.、Singh, S. 和 Guestrin, C. “我为什么应该相信你？”：解释预测任何分类器。在过程中。第 22 届 ACM SIGKDD 国际。会议。关于知识发现和数据挖掘 1135–1144（ACM，2016）。

74. Manning, C. D.、Clark, K.、Hewitt, J.、Khandelwal, U. 和 Levy, O. 通过自我监督训练的人工神经网络中的新兴语言结构。过程。国家科学院。科学。117, 30046–30054 (2020)。

75. Simonsohn, U.、Nelson, L. D. 和 Simmons, J. P. P 曲线：打开文件抽屉的钥匙。J.Exp. 心理。创 143, 534–547 (2013)。

76. Messick, S. 心理评估的有效性：验证从人的反应和表现中得出的推论，作为对分数意义的科学探究。是。心理。50, 741–749 (1995)。

77. Judd, C. M.、Westfall, J. & Kenny, D. A. 将刺激视为社会心理学中的随机因素：对普遍存在但在很大程度上被忽视的问题的一种新的全面解决方案。J.佩尔斯。苏克。心理。103, 54–69 (2012)。

78. 王, T.等人。同伴咨询的指标：三角测量在线成功结果治疗平台。在过程中。2023 年 CHI 会议计算系统中的人为因素 1-17（ACM，2023）。

79. Nook, E. C.、Hull, T. D.、Nock, M. K. 和 Somerville, L. H. 心理距离的语言测量在大量心理治疗记录中跟踪症状水平和治疗结果。过程。国家科学院。科学。119, e2114737119 (2022)。

80. 沃伊特, R.等人。警察随身摄像机镜头中的语言显示了警察尊重方面的种族差异。过程。国家科学院。科学。114, 6521–6526 (2017)。

81. Paullada, A.、Raji, I. D.、Bender, E. M.、Denton, E. 和 Hanna, A. 数据及其（dis）内容：机器学习研究中数据集开发和使用的调查。模式 2, 100336 (2021)。

82. 王, A.等人。GLUE：自然环境的多任务基准测试和分析平台语言理解。在过程中。2018 EMNLP 工作坊。BlackboxNLP：分析和解释 NLP 353-355 的神经网络（计算语言学协会，2018 年）。

83. 格尔曼, S.等人。GEM 基准：自然语言生成及其评估和指标。在过程中。第一个工作坊。关于自然语言生成、评估和度量（GEM 2021）96-120（计算语言学协会，2021）。

84. Birhane, A. 和 Guest, O. 走向计算科学的非殖民化。预印本位于 arXiv <http://arxiv.org/abs/2009.14258> (2020)。

85. Birhane, A. 算法不公正：一种关系伦理方法。模式 2, 100205 (2021)。

86. Erscoi, L.、Kleinherenbrink, A. 和 Guest, O. 皮格马利翁置换：当人工智能人性化时，女性就变得非人性化。SocArXiv 预印本 <https://doi.org/10.31235/osf.io/jqxb6> (2023)。

87. Guest, O. 和 Martin, A. E. 关于大脑、行为和人工神经网络的逻辑推理。计算。大脑行为。6, 213–227 (2023)。

88. 洛伦斯, A.等人。学术界的性别偏见：一个需要解决的终生问题。神经元 109, 2047–2074 (2021)。

89. Metz, C. 和 Weise, K. 微软在 ChatGPT 的创建者身上下了重注，以争夺人工智能的主导地位。《纽约时报》（2023 年 1 月 12 日）。

90. Tesfagerish, S. G.、Kapočiūtė-Dzikienė, J. 和 Damaševičius, R. 使用句子转换器和集成学习进行半监督情感分析的零样本情感检测。应用。科学。12、8662（2022）。

91. 埃尔谢里夫, M.等人。潜在仇恨：理解隐性仇恨言论的基准。在过程中。2021 年会议关于自然语言处理中的经验方法 345-363（计算语言学协会，2021 年）。

92. 普里赞特, R.等人。自动消除文本中的主观偏见。过程。AAAI 会议阿蒂斯。英特尔。34, 480–489 (2020)。

93. Ophir, Y.、Tikochinski, R.、Asterhan, C. S. C.、Sisso, I. 和 Reichart, R. 深度神经网络从 Facebook 文本帖子中检测自杀风险。科学。代表 10, 16685 (2020)。

94. Basta, C.、Costa-jussà, M. R. 和 Casas, N. 评估潜在的性别偏见上下文化的词嵌入。在过程中。第一个工作。关于自然语言处理中的性别偏见33-39（计算语言学协会，2019）。

95. Ashokkumar, A. 和 Pennebaker, J. W. 社交媒体对话揭示了美国各城市因 COVID-19 的爆发而造成的巨大心理变化。科学。副词。7、eabg7843（2021）。

96. Rathje, S.等人。GPT是多语言心理文本分析的有效工具。PsyArXiv 预印本 <https://psyarxiv.com/sekf5/> (2023)。

97. Seraj, S.、Blackburn, K. G. 和 Pennebaker, J. W. 在社交媒体上留下的语言暴露了浪漫分手的情感和认知成本。过程。国家科学院。科学。118、e2017154118 (2021)。

98. 萨普, M.等人。量化想象与自传故事的叙事流程。过程。国家科学院。科学。119, e2211715119 (2022)。

自然评论心理学 |第 2 卷| 2023 年 11 月| 688–701

700

99. Michelmann, S.、Kumar, M.、Norman, K. A. 和 Toneva, M. 大语言模型可以像人类一样分割叙事事件。预印本位于 arXiv <http://arxiv.org/abs/2301.10297> (2023)。

100. 张, S., 她, J.S., 格斯坦伯格, T. 和 罗斯, D. 你就是你的目的：大型语言模型中的本质主义分类。在过程中。安.见面。认知科学学会卷。 45 (2023) 。

101. Cimpian, A. & Salomon, E. 内在启发式：理解世界的直观手段，以及心理本质主义的潜在先驱。行为。脑科学。 37, 461–480 (2014)。

102. Portelance, E.、Degen, J. 和 Frank, M.C. 使用循环神经网络预测早期单词学习中的习得年龄。在过程中。安.见面。认知科学学会 (2020) 。

103. Westerveld, M. F. & Roberts, J. M. A. 自闭症谱系学龄前儿童的口头叙述理解和生产能力。郎.言语听到。服务。施。 48, 260–272 (2017)。

104. Siddaway, A. P.、Wood, A. M. 和 Hedges, L. V. 如何进行系统评价：进行和报告叙述性评价、荟萃分析和荟萃综合的最佳实践指南。安努。心理学牧师。 70, 747–770 (2019)。

105. Tipton, E.、Pustejovsky, J. E. 和 Ahmadi, H. 心理学、教育和医学元回归的当前实践。资源。合成器。方法。 10, 180–194 (2019)。

106. Aher, G.、Arriaga, R. I. 和 Kalai, A. T. 使用大型语言模型来模拟多个人类并复制人类受试者研究。预印本位于 arXiv <http://arxiv.org/abs/2208.10264> (2023)。

107. Pennycook, G.等人。将注意力转移到准确性可以减少网上的错误信息。自然 592, 590–595 (2021)。

108. Brady, W. J.、Wills, J. A.、Burkart, D.、Jost, J. T. 和 Van Bavel, J. J. 政治领导人之间社交媒体上道德化内容传播的意识形态不对称。J.Exp。心理。创世记 148, 1802–1813 (2019)。

109. Milkman, K.L.等人。大型研究提高了应用行为科学的影响。自然 600, 478–483 (2021)。

110. 斯塔德, E.等人。大型语言模型可以改变行为医疗保健的未来：负责任的开发和评估的提案。预印本位于 PsyArXiv <https://doi.org/10.31234/osf.io/cuzvr> (2023) 。

111. 雅各布斯, J.等人。促进数学课堂上的丰富讨论：使用个性化、自动化的反馈来支持反思和教学变革。教。教。教育。 112, 103631 (2022)。

112. Hunkins, N.、Kelly, S. 和 D’ Mello, S. “干得漂亮，你们是摇滚明星！”：教师分析发现支持或破坏学生动机、身份和行为的话语课堂归属感。在 LAK22 中：第 12 中队。学习分析和知识会议。 230–238 (ACM, 2022) 。

113. 施瓦茨, H.A.等人。DLATK：差异语言分析工具包。在过程中。2017年会议自然语言处理中的经验方法：系统演示 55-60（计算语言学协会，2017 年）。

114. 刘, Y.等人。RoBERTa：一种稳健优化的 BERT 预训练方法。预印本位于 arXiv <https://doi.org/10.48550/arXiv.1907.11692> (2019)。

115. 沃尔顿, G.M.等人。在哪里以及有谁进行短暂的社会归属干预可以促进大学进步？科学 380, 499–505 (2023)。

116. 克拉珀, M.等人。评估LLM在中学数学中成长心态支持语言的生成。人工智能在教育领域的应用。在过程中。工作什。关于教育技术与开发的公平、多样性和包容性第 24 届国际会议。会议。关于教育中的人工智能（2023）。

117. Hecht, C. A.、Yeager, D. S.、Dweck, C. S. 和 Murphy, M. C. 信念、可供性和青少年发展：十年成长心态干预的经验教训。副词。孩子。开发。行为。 61, 169–197 (2021)。

118. Hecht, C. A.、Dweck, C. S.、Murphy, M. C.、Kroeper, K. M. 和 Yeager, D. S. 有效探索行为科学中情境调节因素的因果作用。过程。国家科学院。科学。 120, e2216315120 (2023)。

致谢

这项工作得到了美国国家科学基金会的资助，资助号为 1761179 和 2201928（PI：D.S.Y.），由美国国立卫生研究院资助，资助号为 R01HD084772（PI：D.S.Y.）和 P2CHD042849（人口研究中心），以及 William 和 Melinda 盖茨基金会荣获 INV-047751 和 INV-004519 奖项（PI：D.S.Y.）。这项工作还得到了雅各布斯基金会向 D.S.Y. 和以人为中心的人工智能研究所提供的高级研究奖学金的支持。从斯坦福大学到 J.C.E.内容完全由作者负责，并不一定代表资助机构的官方观点。作者还感谢 C. Smith 创建了原始提交中包含的图形的原始版本。术语表定义由 GPT-4 于 2023 年 5 月生成并由作者编辑。

作者贡献

主要作者 D.D.、D.Y.和 D.S.Y.。（同等贡献，按字母顺序排列）构思论文，概述并撰写初稿，指导共同创作过程，提供关键编辑，构思和监督图形、方框和表格的创建，并最终确定提交的手稿版本。资深作者 C.S.D.、J.J.G.和 J.W.P.（按字母顺序列出）协助手稿、方框和表格的概述、组织和概念化，并提供多轮关键编辑。J.C.E.协助概述论文，撰写关键部分的初稿并编辑草稿。所有其他作者都协助提供了实证例子、论文中关键论点和结论的概念化，并提供了批判性的编辑。

利益竞争

J.W.P.是 Pennebaker Conglomerates 的首席执行官，这是一家销售自然语言处理软件和服务的公司。S.C. 和 L.L. 是 Google LLC 的员工，该公司拥有 LLM 技术。D.K.-C.曾是 Google LLC 的员工。

附加信息

同行评审信息 Nature Reviews Psychology 感谢 William Brady 和其他匿名审稿人对这项工作的同行评审所做的贡献。

出版商说明施普林格·自然对于已出版地图和机构隶属关系中的管辖权主张保持中立。

Springer Nature 或其许可方（例如协会或其他合作伙伴）根据与作者或其他权利持有者签订的出版协议拥有本文的专有权；作者对本文已接受的手稿版本的自行存档仅受此类出版协议和适用法律的条款的约束。

© 施普林格自然美国公司 2023