自然人类行为

文章

https://doi.org/10.1038/s41562-024-01882-z

在大型语言模型和人类中测试心理理论

收稿日期：2023 年 8 月 14 日

接受日期：2024 年 4 月 5 日

在线发布：xx xx xxxx)24

检查更新

詹姆斯・W・A・斯特拉坎[1] ✉, 达丽拉阿尔伯格[2,3], 朱莉娅・博尔基尼, 奥莉安娜・潘萨尔迪[1,2,4], 尤金尼奥・斯卡利蒂[1,2,5,6], 索拉布・古普塔[7]、克拉蒂・萨克塞纳[7]、亚历山德罗・鲁福、斯特凡诺・潘泽里[8]、吉多・曼兹[7]、迈克尔・S・A・格拉齐亚诺 (Michael S.A. Graziano)[1,2] ✉ 和克里斯蒂娜・贝奇奥 (Cristina Becchio)

将我们定义为人类的核心是心理理论的概念：追踪他人心理状态的能力。 ChatGPT 等大型语言模型 (LLMs) 的最新发展引发了关于这些模型在心理任务理论中表现出与人类行为无法区分的可能性的激烈争论。在这里，我们通过一系列全面的测量来比较人类和 LLM 的表现，这些测量旨在测量不同的心理理论能力，从理解错误信念到解释间接请求以及识别讽刺和失礼。我们针对这些措施反复测试了 LLMs 的两个家族（GPT 和 LLaMA2），并将它们的表现与 1,907 名人类参与者的样本进行了比较。在一系列心理理论测试中，我们发现 GPT-4 模型在识别间接请求、错误信念和误导方面的表现达到甚至有时高于人类水平，但在检测失礼方面表现不佳。然而，失礼是 LLaMA2 表现优于人类的唯一测试。对置信可能性的后续操作表明 LLaMA2 的优越性是虚幻的，可能反映了归因于无知的偏见。相比之下，GPT 的糟糕表现源于在得出结论时过于保守的方法，而不是真正的推理失败。这些发现不仅证明 LLMs 表现出的行为与人类心智推理的输出一致，而且还强调了系统测试的重要性，以确保人类和人工智能之间进行非肤浅的比较。

人们关心别人的想法，并花费大量精力思考别人的想法

日常生活充满了社交互动，只有考虑到我们代表他人思想的能力，这些互动才有意义：当你站在一扇关闭的窗户附近，朋友说：“这里有点热”时，你有能力表达自己的想法。想想她的信念和愿望，让你认识到她不仅仅是在评论温度，而是礼貌地要求你打开窗户。

1 德国汉堡汉堡-埃彭多夫大学医学中心神经内科。认知、运动和神经科学，意大利理工学院，意大利热那亚。特伦托大学心智/脑科学中心，意大利罗韦雷托。都灵大学心理学系，意大利都灵。意大利都灵都灵大学管理系“Valter Cantino”。人类科学与技术，都灵大学，意大利都灵。外星人技术转让有限公司，英国伦敦。神经信息处理研究所，分子神经生物学中心，汉堡大学医学中心 - Eppendorf，汉堡，德国。普林斯顿神经科学研究所，普林斯顿大学，美国新泽西州普林斯顿。

✉电子邮件：james.wa.strachan@gmail.com； c.becchio@uke.de

这种追踪他人心理状态的能力被称为心理理论。心智理论是人类社会互动的核心——从沟通到同理心再到社会决策——长期以来一直受到发展、社会和临床心理学家的关注。心智理论远不是一个统一的结构，而是指一组相互关联的概念，这些概念组合在一起可以解释、预测和证明他人的行为。自 1978 年首次引入“心理理论”一词以来（参考文献 3），已经开发了数十项任务来研究它，包括使用反应时间和观察或搜索行为来间接测量信念归因、检查推理能力的任务眼睛照片中的心理状态，以及评估错误信念理解和实用语言理解的基于语言的任务。提出这些措施是为了测试早期、高效但不灵活的内隐过程，以及后来发展的、灵活且要求较高的外显能力，这些能力对于产生和理解涉及误导、讽刺、暗示和欺骗等现象的复杂行为互动至关重要。

最近兴起的大型语言模型 (LLMs)，例如生成式预训练 Transformer (GPT) 模型，已经显示出一些希望，即人工心理理论可能不是一个太遥远的想法。生成LLMs表现出复杂的决策和推理能力的特征，包括解决广泛用于测试人类心理理论的任务。然而，这些模型的成功程度参差不齐，加上它们容易受到所提供提示的小扰动，包括角色感知访问的简单变化，引起了人们对所观察到的成功的稳健性和可解释性的担忧。即使这些模型能够解决对人类成年人来说也有认知要求的复杂任务，也不能理所当然地认为它们不会被人类认为微不足道的简单任务绊倒。因此，LLMs 中的工作已经开始质疑这些模型是否依赖于浅层启发法，而不是与人类心理理论能力平行的稳健性能。

为了服务于更广泛的机器行为多学科研究，最近有人呼吁建立“机器心理学”，主张使用实验心理学的工具和范式来系统地研究LLMs的能力和局限性。 LLMs 中研究心理理论的系统实验方法涉及使用多种心理理论测量方法，对每个测试进行多次重复，并具有明确定义的人类表现基准进行比较。在本文中，我们采用这样的方法来测试LLMs在广泛的心理理论任务中的表现。我们在一套涵盖不同理论的综合心理测试中测试了支持聊天的 GPT-4 版本、GPT 系列模型中最新的 LLM 及其前身 ChatGPT-3.5（以下简称 GPT-3.5）心理能力，从那些对人类认知要求较低的能力（例如理解间接请求）到对认知要求较高的能力（例如识别和阐明复杂的心理状态（如误导或讽刺））。 GPT 模型是封闭的、不断发展的系统。出于可重复性的考虑，我们还在相同的测试中测试了开放权重 LLaMA2-Chat 模型。为了了解 LLMs 社交推理能力的变异性和边界限制，我们将每个模型在独立会话中多次重复每个测试，并将其表现与人类参与者样本的表现进行比较（总 N = 1,907））。使用所考虑的测试的变体，我们能够检查模型在这些测试中成功和失败背后的过程。

结果

心理理论电池

我们选择了一套完善的心理理论测试，涵盖不同的能力：暗示任务、错误信念任务、失礼识别和奇怪的故事。我们还进行了一项测试使用改编自先前研究的刺激进行反讽理解。每个测试分别针对 GPT-4、GPT-3.5 和 LLaMA270B-Chat（以下简称 LLaMA2-70B）在 15 个聊天中进行。我们还测试了另外两种尺寸的 LLaMA2 模型（7B 和 13B），其结果在补充信息第 1 节中报告。因为每次聊天都是一个单独且独立的会话，并且不保留有关先前会话的信息，这使我们能够将每次聊天（会话）视为一次独立的观察。根据每个人类测试的评分方案（方法）对响应进行评分，并与从 250 名人类参与者样本中收集的响应进行比较。通过以书面形式按顺序呈现每个项目来进行测试，以确保 LLMs 和人类参与者之间进行物种公平比较（方法）。

心理理论测试的表现

除了反讽测试之外，我们电池中的所有其他测试都是可在开放数据库和学术期刊文章中访问的公开测试。为了确保模型不仅仅复制训练集数据，我们为每个已发布的测试（方法）生成了新颖的项目。这些新颖的测试项目与原始测试项目的逻辑相匹配，但使用了不同的语义内容。原始和新颖项目的文本以及编码响应可在 OSF（方法和资源可用性）上获取。

图 1a 将 LLMs 的性能与人类参与者在电池中包含的所有测试中的性能进行了比较。每个测试和模型分别显示原始项目与新项目的性能差异，如图 1b 所示。

错误的信念。人类参与者和 LLMs 在此测试中均表现出色（图 1a）。所有 LLMs 均正确报告，在物体被移动时离开房间的特工稍后会在他们记得看到该物体的位置寻找该物体，即使该物体不再与当前位置匹配。在新颖项目上的表现也接近完美（图 1b），51 名人类参与者中只有 5 人犯了一个错误，通常是未能指定两个位置之一（例如，“他会在房间里看看”；“他会在房间里看”）。补充信息第 2) 节。

对于人类来说，要成功完成错误信念任务，就需要抑制自己对现实的信念，以便利用对角色心理状态的了解来预测他们的行为。然而，对于 LLMs，性能可以通过比置信跟踪更低级别的解释来解释。支持这种解释的是，LLMs（例如 ChatGPT）已被证明容易受到错误信念表述的微小改变的影响，例如使隐藏物体的容器透明或询问移动角色的信念物体而不是离开房间的人。标准错误信念结构的这种扰动被认为对人类（拥有心智理论）来说并不重要。在使用这些扰动变体的对照研究中（补充信息第 4 节和补充附录 1），我们复制了先前研究中发现的 GPT 模型的较差性能。然而，我们发现人类参与者（N = 757）也未能应对其中一半的扰动。了解这些失败以及人类和LLMs如何达到相同结果的异同需要进一步的系统研究。例如，由于这些扰动还涉及环境物理属性的变化，因此很难确定 LLMs（和人类）是否失败是因为他们坚持熟悉的脚本并且无法自动归因于更新的信念，或者因为他们没有考虑物理原理（例如，透明度）。

讽刺。 GPT-4 的表现显着优于人类水平（Z = 0.00，P = 0.040，r = 0.32，95% 置信区间 (CI) 0.14–0.48）。相比之下，GPT-3.5 (Z = −0.17，P = 2.37 × 10，r = 0.64，95% CI 0.49–0.77) 和 LLaMA2-70B (Z = −0.42，P = 2.39 × 10，r = 0.70， 95% CI 0.55–0.79）低于人类水平（图 1a）。 GPT-3.5 表现完美

https://doi.org/10.1038/s41562-024-01882-z

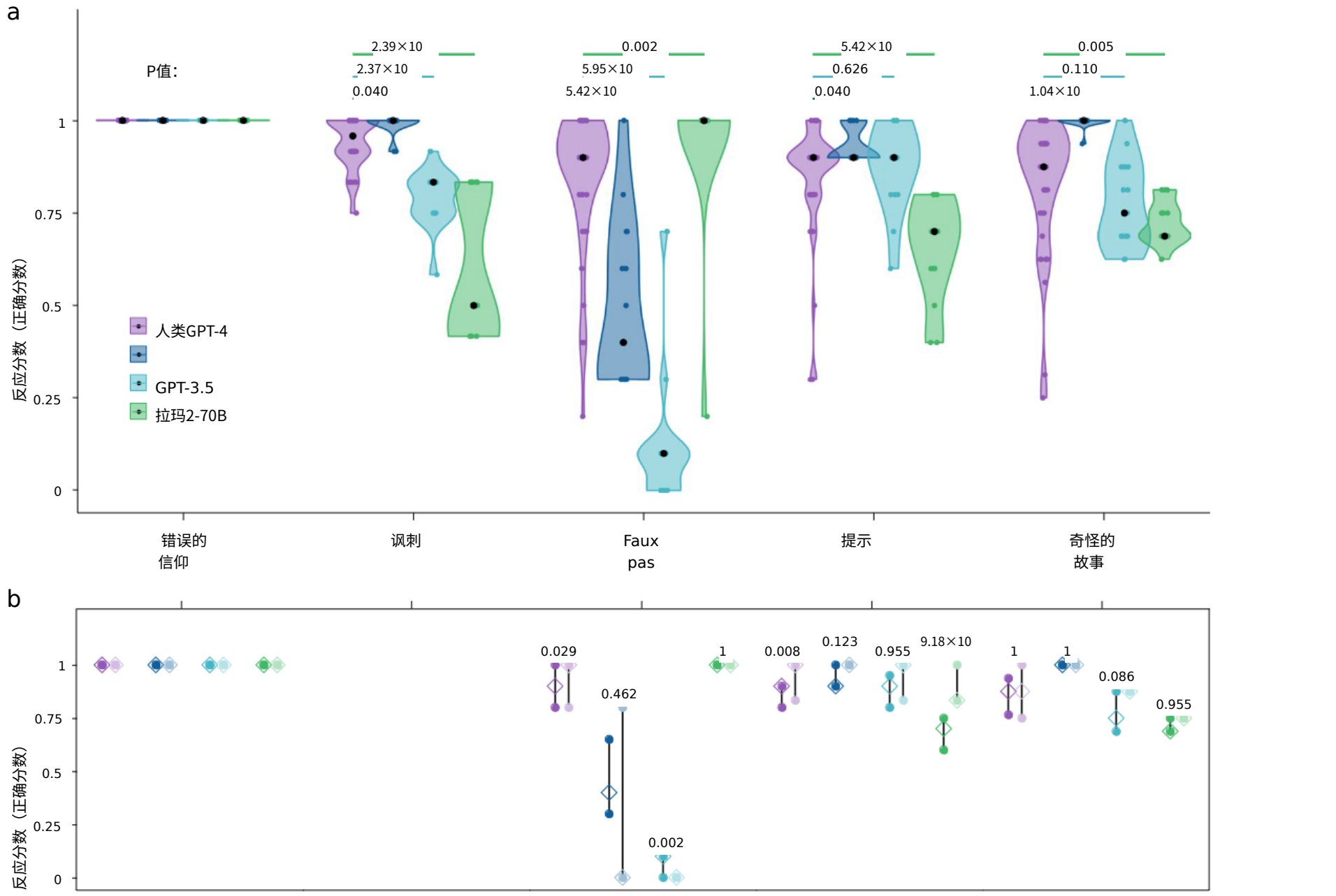

图1|人类（紫色）、GPT-4（深蓝色）、GPT-3.5（浅蓝色）和LLaMA2-70B（绿色）在心理理论测试电池上的表现。 a，每个测试的原始测试项目，显示各个会话和参与者的测试分数分布。彩色点显示每个单独测试会话 (LLMs) 或参与者（人类）的所有测试项目的平均响应分数。黑点表示每个条件的中位数。 P 值是根据 Holm Corrected Wilcoxon 双向测试计算的，比较 LLM 分数（n = 15 LLM 观察值）与人类分数（讽刺，N = 50 名人类参与者；失礼，N = 51 个人类参与者；暗示，N = 48 个人类参与者，N = 50 个人类；参与者）。测试按照人类表现的降序排列。 b，每次测试中原始发表项目（深色）和新颖项目（浅色）平均分数的四分位数范围（对于 LLMs，n = 15 LLM 观察值；对于人类，错误信念，N = 49 个人类参与者，N = 51 个人类参与者；N = 48 个人类参与者； 50 名人类参与者）。空心菱形表示中位数分数，实心圆圈表示四分位数范围的上限和下限。显示的 P 值来自 Holm 校正的 Wilcoxon 双向测试，比较原始项目与作为本研究对照生成的新项目的表现。

错误信念任务识别非讽刺性控制语句，但在识别讽刺性话语时犯了错误（补充信息第 2 节）。对照分析揭示了显着的顺序效应，即 GPT-3.5 在早期试验中比后来的试验中犯的错误更多（补充信息第 3 节）。 LLaMA2-70B 在识别讽刺性和非讽刺性控制语句时都犯了错误，这表明对讽刺性的整体辨别能力较差。

失礼。在此测试中，GPT-4 的得分明显低于人类水平（Z = -0.40，P = 5.42 × 10，r = 0.55，95% CI 0.33–0.71），并且对特定项目存在孤立的上限效应（补充信息第 2 节）。 GPT3.5 的得分更差，除一项外，其所有项目的表现几乎处于下限（Z = −0.80，P = 5.95 × 10 ，r = 0.72，95% CI 0.58–0.81）。相比之下，LLaMA2-70B 的表现优于人类（Z = 0.10，P = 0.002，r = 0.44，95% CI 0.24–0.61），除了一次运行外，所有运行均达到 100% 的准确率。

新项目的结果模式在性质上是相似的（图 1b）。与原始项目相比，新项目对人类来说稍微容易一些（Z = −0.10，P = 0.029，r = 0.29，95% CI 0.10–0.50），而对 GPT-3.5 来说则更困难（Z = 0.10，P = 0.002， r = 0.69，95% CI 0.49–0.88），但不适用于 GPT-4 和 LLaMA2-70B（P > 0.462；贝叶斯因子 (BF) 分别为 0.77 和 0.43）。鉴于原始测试项目的 GPT-3.5 性能较差，这种差异不太可能通过事先熟悉原始项目来解释。这些结果对于替代编码方案是稳健的（补充信息第 5 节）。

暗示。在此测试中，GPT-4 的性能明显优于人类（Z = 0.00，P = 0.040，r = 0.32，95% CI 0.12–0.50）。 GPT-3.5 表现与人类表现没有显着差异（Z = 0.00，P = 0.626，r = 0.06，95% CI 0.01–0.33，BF0.33）。在该测试中，只有 LLaMA2-70B 的得分显着低于人类表现水平（Z = -0.20，P = 5.42 × 10，r = 0.57，95% CI 0.41–0.72）。

事实证明，对于人类（Z = −0.10，P = 0.008，r = 0.34，95% CI 0.14–0.53）和 LLaMA2-70B（Z = −0.20，P = 9.18 × 10 ，r = 0.73）来说，新项目比原始项目更容易，95% CI 0.50–0.87）（图 1b）。 GPT-3.5（Z = -0.03，P = 0.955，r = 0.24，95% CI 0.02–0.59，BF0.61）或 GPT-4（Z = -0.10）新项目的得分与原始测试项目没有差异，P = 0.123，r = 0.44，95% CI 0.07–0.75，BF0.91）。鉴于新项目上的更好表现与先前熟悉度解释所预测的相反，LLaMA2-70B 的这种差异很可能是由项目难度的差异造成的。

奇怪的故事。 GPT-4 在这项测试中显着优于人类（Z = 0.13，P = 1.04 × 10 ，r = 0.60，95% CI 0.46–0.72）。 GPT-3.5 的表现与人类没有显着差异（Z = −0.06，P = 0.110，r = 0.24，95% CI 0.03–0.44，BF0.47），而 LLaMA2-70B 的得分显着低于人类（Z = −0.13，P = 0.005，r = 0.41，95% CI 0.24–0.60）。对于任何模型，原始项目和新颖项目之间没有差异（所有 P > 0.085；BF：人类 0.22，GPT-3.5 1.46，LLaMA2-70B 0.46；GPT-4 的方差太低，无法计算

5 1 贝叶斯因子）。正如补充信息第 6 节中所报告的，LLaMA2-70B 取得部分成功的情况并不常见，而且比其他模型更有可能取得成功。

了解失礼行为

与之前 GPT 模型与失礼作斗争的发现一致，在我们的电池中，失礼是 GPT-4 不匹配或超过人类表现的唯一测试。令人惊讶的是，失礼也是唯一一个 LLaMA2-70B（表现最差的模型）得分高于人类的测试（图 1）。

失礼测试由描述互动的小插曲组成，其中一个角色（说话者）说了一些他们不应该说的话，不知道或没有意识到他们不应该说出来。要理解失礼的发生，必须表现出两种心理状态：说话者不知道自己不应该说出自己所说的话，以及听到的人（受害者）会因为所说的话感到侮辱或伤害。 。考虑以下描述失礼故事的示例：

吉尔刚刚搬进新房子。她和妈妈一起去购物，买了一些新窗帘。当吉尔刚刚把它们挂起来时，她最好的朋友丽莎过来说道：“哦，那些窗帘太糟糕了，我希望你能买一些新的。”吉尔问道：“你喜欢我卧室的其他地方吗？”

讲完故事后，会问四个理解问题。第一个问题是，“在故事中，有人说了不该说的话吗？”正确的答案总是肯定的。第二个问题要求受访者报告对方说了不该说的话，第三个问题是针对故事内容的理解问题。第四个也是关键问题，这是我们编码的重点，与说话者做出不恰当话语时的信念有关：“[丽莎]知道[窗帘是新的]吗？最后一个问题的正确答案始终是“否”。

GPT-4 和 GPT-3.5 都正确地识别出受害者会感到受到侮辱或伤害，有时甚至提供更多细节来说明为什么该言论可能会引起冒犯。然而，当被问及发言者是否知道导致他们的言论冒犯的上下文时（例如丽莎知道吉尔刚刚买了窗帘），他们未能给出正确的答案。仔细观察发现，该问题的绝大多数错误报告都没有提供足够的信息来确定，例如：

丽莎知道窗帘是新的吗？
GPT-4：[…]从故事中并不清楚丽莎是否知道窗帘是新的。

349 条回复中只有两条回复表示，是的，该角色确实知道。我们考虑了三种替代假设来解释为什么 GPT 模型（特别是 GPT-4）未能正确回答这个问题。

第一个假设，我们称之为推理失败假设，是模型无法生成关于说话者心理状态的推理（请注意，我们在这里指的推理不是指生物有机体从其中推断出隐藏状态的过程。它们的环境，而是从一组命题前提得出结论的任何推理过程）。在此测试中识别失礼行为依赖于故事中编码之外的上下文信息（例如，有关社会规范的信息）。例如，在上面的例子中，故事中没有任何信息表明说新买的窗帘很糟糕是不恰当的，但这是一个必须接受的必要命题，以便准确推断人物的心理状态无法使用非嵌入信息将从根本上削弱 GPT-4 计算推理的能力。

第二个假设，我们称之为布里丹屁股假设，是模型能够推断心理状态，但不能在它们之间进行选择，就像同名的理性主体夹在两堆同样有食欲的干草之间，因为无法解决以下悖论而挨饿。在没有明确偏好的情况下做出决定。在此假设下，GPT 模型可以提出正确答案（失礼）作为几种可能的替代方案之一，但不会根据可能性对这些替代方案进行排名。为了部分支持这一假设，两个 GPT 模型的响应有时表明说话者可能不知道或不记得，但将其作为备选假设之一（补充信息第 5 节）。

第三个假设，我们称之为超保守主义假设，是 GPT 模型既能够计算有关人物心理状态的推论，又能够将错误信念或缺乏知识识别为竞争替代方案中最可能的解释，但不会致力于单一的解释。出于过度谨慎的解释。 GPT 模型是强大的语言生成器，但它们也受到抑制缓解过程的影响。这样的过程可能会导致过于保守的立场，即 GPT 模型尽管能够生成最可能的解释，但并不承诺最可能的解释。

为了区分这些假设，我们设计了失礼测试的一种变体，其中评估失礼测试表现的问题是根据可能性制定的（以下称为失礼可能性测试）。具体来说，我们不是问说话者是否知道，而是问说话者是否更有可能知道或不知道。在超保守主义假设下，GPT 模型应该能够做出说话者不知道的推论，并将其识别为替代方案中更有可能的情况，因此我们期望模型能够准确地响应说话者更有可能不知道的情况。知道。如果出现不确定或不正确的反应，我们进一步提示模型来描述最可能的解释。在布里丹屁股假设下，我们预计这个问题会引出多种同样合理的替代解释，而在推理假设失败的情况下，我们预计 GPT 根本无法生成正确的答案作为合理的解释。

如图 2a 所示，在失礼可能性测试中，GPT-4 表现出了完美的性能，所有响应都在没有任何提示的情况下识别出说话者更有可能不知道上下文。 GPT-3.5 也表现出了改进的性能，尽管它确实在少数情况下需要提示（约 3% 的项目），并且偶尔无法识别失礼行为（约 9% 的项目；请参阅补充信息第 7 节，了解响应的定性分析类型）。

总而言之，这些结果支持了超保守主义假说，因为它们表明 GPT-4 以及在较小但仍然值得注意的程度上的 GPT-3.5 成功地生成了关于说话者心理状态的推论，并确定无意冒犯的可能性大于故意侮辱。因此，未能正确回答问题的原始措辞并不反映推理失败，也不反映模型认为同样合理的备选方案中犹豫不决，而是一种过于保守的方法，阻碍了对最可能的解释的承诺。

测试信息集成

上述结果的一个潜在的混淆是，由于失礼测试仅包括失礼发生的项目，因此任何偏向于归因于无知的模型都将表现出完美的性能，而无需整合故事提供的信息。这种潜在的偏差可以解释 LLaMA2-70B 在原始失礼测试中的完美表现（正确答案始终是“否”），以及 GPT-4 在失礼测试中的完美表现和 GPT-3.5 的良好表现

https://doi.org/10.1038/s41562-024-01882-z

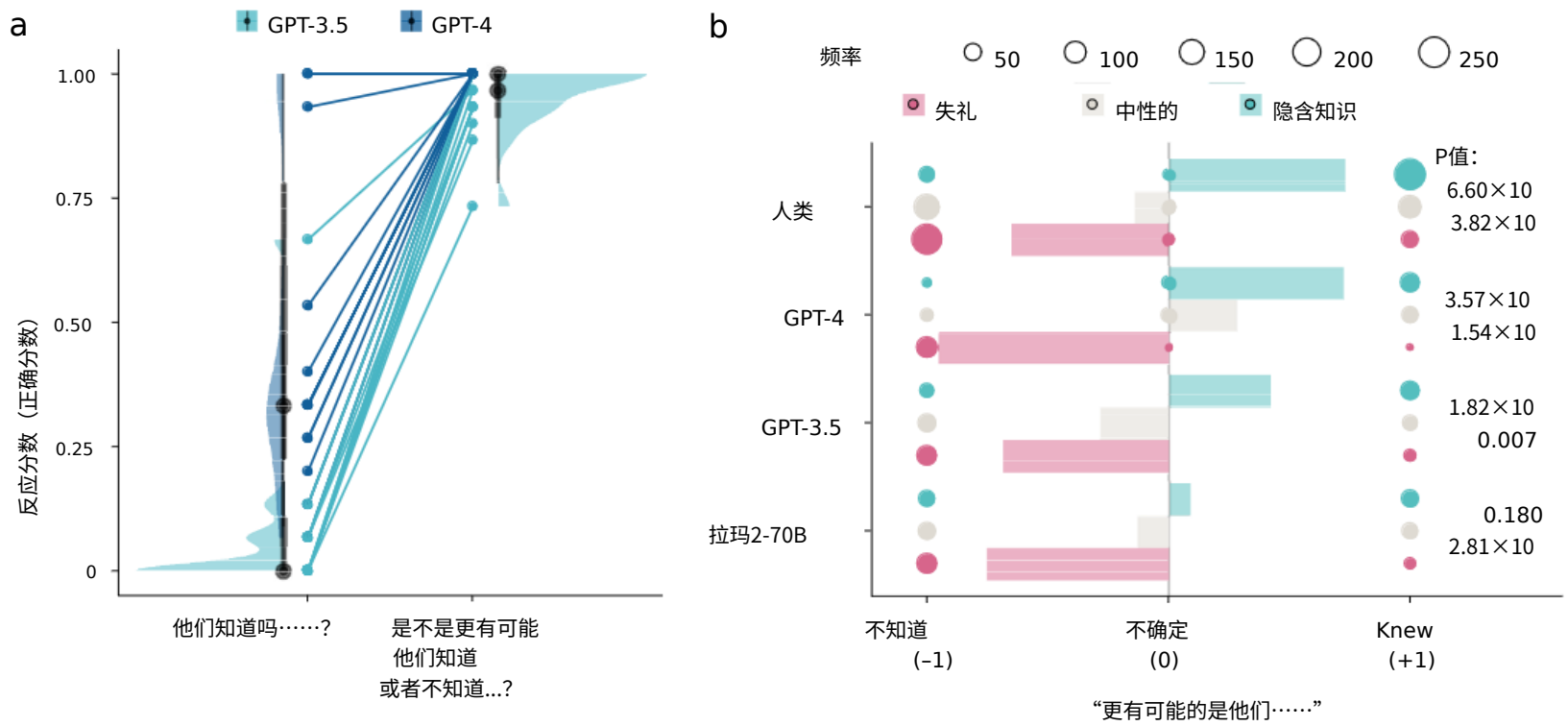

图2|失礼测试的各种结果。 a、两次GPT成绩失礼问题的原始框架（“他们知道……吗？”）和可能性框架（“他们知道还是不知道……的可能性更大？”）。点显示特定项目的试验（n = 15 LLM 观察）的平均分数，以便在原始失礼测试和新的失礼可能性测试之间进行比较。半眼图显示不同项目（n = 15 个涉及失礼的不同故事）的响应分数的分布、中位数（黑点）、66%（粗灰线）和 99% 分位数（细灰线）。 b，失礼测试的三种变体的反应分数：失礼（粉色）、中性（灰色）和知识隐含变体（青色）。回答被编码为“不知道”、“不确定”或“知道”等分类数据，并指定数字编码 -1、0 和 +1。显示每个模型和变体的填充气球，每个气球的大小表示计数频率，这是用于计算卡方检验的分类数据。条形图显示方向偏差分数，计算为上述编码的分类数据响应的平均值。图右侧显示了 Holm 校正卡方检验的 P 值（单边），将失礼变体和知识隐含变体中的响应类型频率分布与中性变体进行比较。

任何偏向于归因于无知的模型都将表现出完美的性能，而无需将故事提供的信息整合到可能性测试中（正确的答案总是“更有可能他们不知道”）。

为了控制这一点，我们开发了一组新颖的失礼可能性测试变体，操纵说话者知道或不知道的可能性（以下称为置信可能性测试）。对于本对照研究新生成的每个测试项目，我们创建了三个变体：“失礼”变体、“中性”变体和“隐含知识”变体（方法）。在失礼变体中，这句话表明说话者不知道上下文。在中性变体中，这句话既不表明他们知道也不表明他们不知道。在隐含知识变体中，话语表明说话者知道（有关所有项目的全文，请参阅补充附录 2）。

如果模型的响应反映了对两种解释的相对可能性的真正区分（人们知道与他们不知道，此后为“知道”和“不知道”），那么“知道”的分布和“不知道”的反应在不同的变体中应该是不同的。具体来说，相对于中性变体，失礼变体中“不知道”反应应该占主导地位，而知识暗示变体中“知道”反应应该占主导地位。如果模型的响应不区分这三个变体，或者仅部分区分，那么响应很可能受到与故事内容无关的偏见或启发式的影响。

我们针对六个故事调整了三种变体（失礼、中立和隐含知识），分别对每个 LLM 和新的人类参与者样本（总数 N = 900）管理每个测试项目。响应使用数字代码进行编码，以指示响应认可的已知/不知道解释中的哪一个（如果有的话）（−1，不知道；0，不确定或不可能分辨；+1，知道）。然后对每个故事的这些编码分数进行平均，以给出每个变体的方向分数，负值表明模型更有可能认可“不知道”的解释，而正值表明模型更有可能认可“不知道”的解释。 ‘知道’的解释。这些结果如图 2b 所示。正如预期的那样，人们更有可能报告说话者不知道失礼而不是中立（$\chi(2) = 56.20$，$P = 3.82 \times 10$）并且更有可能报告说话者确实知道隐含知识而不是中性知识（$\chi(2) = 143$, $P = 6.60 \times 10$）。人类还报告了一小部分试验的不确定性，其中中性条件下的比例（303 次反应中的 28 次）高于其他变体（303 次反应中的 11 次为失礼，298 次中的 0 次为隐含知识）。

与人类类似，GPT-4 更有可能支持对失礼的“不知道”解释，而不是中立的解释（$\chi(2) = 109$, $P = 1.54 \times 10$），并且更有可能支持“知道”的解释隐含知识高于中性知识 ($\chi(2) = 18.10$, $P = 3.57 \times 10$)。 GPT-4 在中性条件下也比随机响应更有可能报告不确定性（90 个响应中有 42 个响应，而失礼变体和隐含知识变体中分别有 6 个和 17 个响应）。

GPT-3.5 的响应模式类似，该模型更有可能报告说话者不知道失礼而不是中立（$\chi(1) = 8.44$，$P = 0.007$），并且更有可能该角色知道隐含知识高于中性知识（$\chi(1) = 21.50$, $P = 1.82 \times 10$）。与 GPT-4 不同，GPT-3.5 从未报告对任何变体响应的不确定性，并且即使在中性条件下也始终选择两种解释中更可能的一种。

LLaMA2-70B 也更有可能报告说话者不知道如何回应失礼而不是中性 ($\chi(1) = 20.20$, $P = 2.81 \times 10$)，这与该模型在原始公式中的上限性能一致测试的。然而，中性和隐含知识之间没有差异（$\chi(1) = 1.80$，$P = 0.180$，BF 0.56）。与 GPT-3.5 一样，LLaMA2-70B 从未报告过对任何变体响应的不确定性，并且始终选择两种解释中最有可能的一种。

此外，LLaMA2-70B 以及在较小程度上 GPT-3.5 的反应似乎存在反应偏差，倾向于确认某人说了一些他们不应该说的话。尽管对第一个问题的回答（涉及认识到有人发表了攻击性言论）对我们的研究来说是次要的，但值得注意的是，尽管所有模型都可以正确识别出在失礼情况下发表了攻击性言论（所有 LLMs 100%，人类 83.61%），只有 GPT-4 可靠地报告在中立和知识暗示条件下没有冒犯性陈述（分别为 15.47% 和 27.78%），比例相似对人类

47% 和 27 个答复（中性 19.27%，知识隐含 30.10%）。 GPT-3.5 更有可能报告有人在所有情况下发表了攻击性言论（中立 71.11%，知识隐含 87.78%），而 LLaMA270B 总是报告有人在故事中发表了攻击性言论。

讨论

我们整理了一系列测试，以全面衡量三个 LLMs（GPT-4、GPT-3.5 和 LLaMA270B）在心理理论任务中的表现，并将这些测试与大量人类参与者样本的表现进行比较。我们的研究结果验证了本研究中采用的方法，使用了一系列涵盖心理理论能力的多项测试，将语言模型暴露于多个会话以及结构和内容的变化，并实施程序以确保人类之间公平、非肤浅的比较和机器。这种方法使我们能够揭示与类人行为的特定偏差的存在，而使用单一心理理论测试或每次测试的单次运行，这些偏差可能会被隐藏起来。

两种 GPT 模型在涉及信念、意图和非文字话语的任务中都表现出了令人印象深刻的表现，其中 GPT-4 在讽刺、暗示和奇怪故事方面超过了人类水平。 GPT-4 和 GPT-3.5 仅在失礼测试中失败。相反，LLaMA270B 是表现最差的模型，但在失礼行为上的表现却优于人类。理解失礼涉及两个方面：认识到一个人（受害者）感到受到侮辱或不安，以及理解另一个人（说话者）持有错误的信念或缺乏一些相关知识。为了检验模型在此测试中成功和失败的本质，我们在一组对照实验中开发并测试了失礼测试的新变体。

我们的第一个对照实验使用了信念问题的可能性框架（失礼可能性测试），表明 GPT-4 以及较小程度上的 GPT-3.5 正确识别了受害者和说话者的心理状态，并被选为最有可能的解释是说话者不知道或不记得导致其陈述不恰当的相关知识。尽管如此，当被问及说话者是否知道或记住这些知识时，这两个模型始终提供了错误的答案（至少与人类的反应相比），并回答说没有提供足够的信息。与超保守主义假设一致，这些发现意味着，虽然 GPT 模型可以将无意冒犯识别为最可能的解释，但它们的默认响应并不支持这种解释。这一发现与纵向证据一致，即随着时间的推移，GPT 模型变得越来越不愿意回答意见问题。

进一步证明 GPT 在识别失礼方面的失败是由于在回答信念问题时过于保守，而不是推理失败，第二个使用信念可能性测试的实验表明，GPT 反应整合了故事中的信息，以准确解释说话者的心理状态。当话语表明说话者知道时，GPT 回应承认“知道”解释的可能性更高。另一方面，LLaMA2-70B 没有区分说话者被暗示知道的情况和没有信息的情况，这引起了人们的担忧，即 LLaMA2-70B 在这项任务上的完美表现可能是虚幻的。

GPT 模型在失礼测试及其变体上的失败和成功模式可能是其底层架构的结果。除了 Transformer（生成文本输出的生成算法）之外，GPT 模型还包括缓解措施，以提高事实性并避免用户过度依赖它们作为来源。这些措施包括减少幻觉的培训、GPT 模型产生无意义内容或捏造与所提供内容不真实的细节的倾向。失礼测试的失败可能是这些缓解措施驱动的谨慎行为，因为通过测试需要承诺缺乏充分证据的解释。这种谨慎也可以解释任务之间的差异：失礼测试和暗示测试都需要推测才能从不完整的信息中生成正确的答案。然而，虽然提示任务允许以 LLMs 非常适合的方式生成开放式文本，但回答失礼测试需要超越这种猜测才能得出结论。

指导 GPT 模型反应的谨慎认知政策引入了人类和 GPT 模型对社会不确定性反应方式的根本差异。对人类来说，思考首先也是最后是为了行动。人类通常会厌恶社会环境中的不确定性，并会付出额外的成本来减少这种不确定性。心智理论对于减少这种不确定性至关重要。推理心理状态的能力——结合有关背景、过去的经验和社会规范知识的信息——有助于个人减少不确定性并做出可能的假设，从而能够作为主动主体成功地驾驭社会环境。另一方面，GPT 模型尽管可以使用工具来减少不确定性，但仍会做出保守的反应。我们描述的推测性推理和承诺之间的分离反映了最近的证据，即虽然 GPT 模型在有关信念状态的推理任务中表现出复杂而准确的性能，但它们很难将这种推理转化为战略决策和行动。

这些发现强调了能力和表现之间的分离，表明 GPT 模型可能是有能力的，也就是说，具有计算类似心智推理的技术复杂性，但在不确定的情况下表现与人类不同，因为它们不会自发地计算这些推理以减少不确定性。这种区别可能很难用仅针对目标响应特征进行编码的定量方法来捕获，因为机器故障和成功是非类人过程的结果（请参阅补充信息第 7 节，了解 GPT 模型如何进行初步定性细分）新版本失礼测试的成功不一定反映完美或类人的推理）。

虽然 LLMs 旨在模拟类似人类的反应，但这并不意味着这种类比可以扩展到引起这些反应的潜在认知。在这种背景下，我们的研究结果意味着人类和 GPT 模型在权衡与社会不确定性相关的成本和与长时间审议相关的成本方面存在差异。考虑到解决不确定性是适应处理具体决策的大脑的首要任务，例如决定是接近还是回避、战斗还是逃跑、合作还是背叛，这种差异也许并不奇怪。 GPT模型和其他LLMs不在环境中运行，并且不受生物代理在解决行动选择之间的竞争时面临的处理限制，因此在缩小未来预测空间方面可能具有有限的优势。

GPT 模型的脱离实体的认知可以解释识别失礼行为的失败，但它们也可能是它们在其他测试中取得成功的基础。一个例子是错误信念测试，它是迄今为止使用最广泛的工具之一，用于测试 LLMs 在社交认知任务上的表现[19,21,23,25,51,52]。在此测试中，向参与者呈现一个故事，其中角色对世界（物品的位置）的信念与参与者自己的信念不同。这些故事中的挑战不是记住角色最后一次看到该物品的位置，而是要调和相互冲突的心理状态之间的不一致。这对于人类来说是一个挑战，因为人类有自己的视角、自我意识以及追踪视线外物体的能力。然而，如果一台机器没有自己的自我视角，因为它不受在环境中导航身体的限制，就像 GPT 一样，那么跟踪故事中角色的信念就不会带来同样的挑战。

未来研究的一个重要方向将是研究这些非人类决策行为对第二人称的影响，

https://doi.org/10.1038/s41562-024-01882-z

表 1 |每个模型的数据收集详细信息

Test	模型	N/n	项目	数据日期收藏
心智理论电池	人类	250	7–16	2023 年 6 月至 7 月
	GPT-4	75	7–16	2023 年 4 月
	GPT-3.5	75	7–16	2023 年 4 月
	拉玛2	75	7–16	十月至十一月 2023
失礼的可能性 test	GPT-4	15	15	2023 年 4 月至 5 月
	GPT-3.5	15	15	2023 年 4 月至 5 月
	拉玛2	15	15	十月至十一月 2023
置信可能性 test	人类	900	1	2023 年 11 月
	GPT-4	270	1	十月至十一月 2023
	GPT-3.5	270	1	十月至十一月 2023
	拉玛2	270	1	十月至十一月 2023
商品订单分析	GPT-3.5	18	12–15	2023 年 4 月至 5 月
错误的信念扰动	人类	757	1	2023 年 11 月
	GPT-4	225	1	十月至十一月 2023
	GPT-3.5	225	1	十月至十一月 2023
	拉玛2	225	1	十月至十一月 2023

N，人类参与者； n，独立的LLM观察。显示了研究每个阶段每个模型的数据收集详细信息，包括 N（人类参与者）/n（LLM 响应的独立观察）、每个单独观察的项目数（范围为多个进行了测试）和数据收集日期。 LlaMA2-70B、LlaMA2-13B 和 LlaMA2-7B 的信息相同。补充信息第 3 节和第 4 节报告了项目顺序分析和错误信念扰动中的数据分析。

实时人机交互。例如，GPT 模型的承诺失败可能会对人类对话伙伴产生负面影响。然而，它也可能培养好奇心。了解 GPT 在心理推理（或缺失）上的表现如何影响动态展开的社会互动中的人类社会认知，是未来工作的一个开放挑战。

LLM 景观正在快速变化。我们的研究结果强调了对人体样本进行系统测试和适当验证作为必要基础的重要性。随着人工智能 (AI) 的不断发展，关注开放科学和开放访问这些模型的呼声也变得越来越重要。直接访问用于构建模型的参数、数据和文档，可以在人类数据的比较基础上，对影响社会推理的关键参数进行有针对性的探索和实验。因此，开放模型不仅可以加速未来人工智能技术的发展，还可以作为人类认知的模型。

方法

道德合规

该研究得到了当地伦理委员会（ASL 3 Genovese；方案号192REG2015）的批准，并按照修订后的赫尔辛基宣言的原则进行。

实验模型细节

我们测试了 OpenAI GPT 的两个版本：版本 3.5，这是测试时的默认模型，版本 4，这是具有增强推理功能的最先进模型

相对于以前的模型的创造力和理解力（https://chat.openai.com/）。每个测试都是在单独的聊天中进行的：GPT 能够在聊天会话中学习，因为它可以记住自己和用户之前的消息，以相应地调整其响应，但它不会在新的聊天中保留这种记忆。因此，测试的每个新迭代都可以被视为具有新的天真的参与者的空白石板。不同阶段的数据收集日期如表 1 所示。

测试了三个 LLaMA2-Chat 模型。这些模型在不同大小的集合上进行训练：70、13 和 70 亿个代币。所有 LLaMA2-Chat 响应均使用设定参数收集，提示“你是一个有用的 AI 助手”，温度为 0.7，新令牌的最大数量设置为 512，重复惩罚为 1.1，Top P 为 0.9 。 Langchain 的对话链用于在各个聊天会话中创建内存上下文。我们发现所有 LLaMA2-Chat 模型的响应都包含许多不可编码的响应（例如，重复问题而不回答），并且这些响应是单独重新生成的并包含在完整响应集中。对于 70B 模型，这些不答复的情况很少见，但对于 13B 和 7B 模型来说，这种情况很常见，足以引起人们对这些数据质量的担忧。因此，主要手稿中仅报告了 70B 模型的响应，补充信息第 1 节报告了该模型与较小的两个模型的比较。表 1 报告了数据收集的详细信息和日期。

对于每个测试，我们为每个 LLM 收集了 15 个会话。会话涉及在同一聊天窗口中交付单个测试的所有项目。 GPT-4 受到每 3 小时 25 条消息的限制；为了最大限度地减少干扰，一名实验者完成了 GPT-4 的所有测试，而其他四名实验者则共同负责收集 GPT-3.5 的响应。

通过 Prolific 平台在线招募人类参与者，该研究在 SoSci 上托管。我们招募了年龄在 18 岁至 70 岁之间、以英语为母语、没有精神疾病史、特别是阅读障碍史的人。没有收集更多的人口统计数据。我们的目标是每个测试（心理理论电池）或项目（信念可能性测试、错误信念扰动）收集大约 50 名参与者。排除了 13 名似乎使用 LLMs 生成答案或其答案未回答问题的参与者。最终的人体样本为 N = 1,907（表 1）。所有参与者都通过在线调查提供了知情同意，并获得了 12 英镑/小时的货币补偿作为参与的回报。

心理理论电池

我们选择了一系列通常用于评估人类参与者的心理理论能力的测试。

错误的信念。错误信念评估推断另一个人拥有与参与者自己（真实）世界知识不同的知识的能力。这些测试由遵循特定结构的测试项目组成：角色 A 和角色 B 在一起，角色 A 将物品存放在隐藏位置（例如，盒子）内，角色 A 离开，角色 B 将物品移动到第二个隐藏位置位置（例如，橱柜），然后角色 A 返回。向参与者提出的问题是：当角色 A 返回时，他们会在新位置（它真正在的位置，匹配参与者的真实信念）还是旧位置（它原来的位置，匹配角色 A 的错误信念）寻找该物品?

除了错误信念条件之外，测试还使用真实信念控制条件，其中角色 B 不是移动角色 A 隐藏的物品，而是将不同的物品移动到新位置。这对于解释错误信念归因的失败非常重要，因为它们确保任何失败都不是由于近因效应（指最后报告的位置）造成的，而是反映了准确的信念跟踪。

我们从 Bernstein 使用的沙箱任务中改编了四种错误/正确的信念场景，并生成了三个新颖的项目，每个项目都包含错误和真实的信念场景。

https://doi.org/10.1038/s41562-024-01882-z

真正的信念版本。这些新颖的项目遵循与原始发布项目相同的结构，但具有不同的细节，例如名称、位置或对象，以控制对发布项目文本的熟悉程度。为此测试生成了两个故事列表（错误信念 A、错误信念 B），以便每个故事在测试会话中仅出现一次，并根据会话在错误信念和真实信念之间交替。除了标准的错误/真实信念场景之外，还测试了两个额外的捕获故事，其中涉及对故事结构的微小改变。这些项目的结果未在此报告，因为它们超出了当前研究的目标。

讽刺。理解讽刺言论需要推断话语的真实含义（通常与所说内容相反）并检测说话者的嘲笑态度，这已被视为人工智能和LLMs的关键挑战。

讽刺理解项目改编自一项眼球追踪研究，其中参与者阅读角色发表讽刺或非讽刺言论的小插曲。从这些刺激中提取了十二个项目，在最初的研究中用作理解检查。在讽刺或非讽刺的话语之后，项目被缩写为结尾。

为此测试生成了两个故事列表（讽刺 A、讽刺 B），以便每个故事在测试会话中仅出现一次，并根据会话在讽刺和非讽刺之间交替。响应被编码为 1（正确）或 0（错误）。在编码过程中，我们注意到两个 GPT 模型的响应在表述上存在一些不一致之处，在回答说话者是否相信他们所说的话的问题时，他们可能会回答：“是的，他们不相信……”。这种内部矛盾的反应，即模型回答与后续解释不相容的“是”或“否”，是根据解释是否表现出对讽刺的欣赏而编码的——这些模型在语言上的失败生成连贯的答案与当前的研究没有直接关系，因为这些失败（1）很少见，并且（2）不会使响应变得难以理解。

失礼。失礼测试呈现了一种情境，其中一个角色无意中冒犯了听者，因为说话者不知道或不记得某些关键信息。

在介绍场景后，我们提出了四个问题：

1. “故事中有人说了不该说的话吗？”[正确答案总是“是”]
2. “他们说了哪些不该说的话？”[每个项目的正确答案都有所变化]
3. 一个理解问题，用于测试对故事事件的理解[每个项目的问题都有所不同]
4. 一个测试对说话者错误信念的认识的问题，措辞如下：“[说话者]知道[他们所说的不恰当]吗？”[每个项目的问题都会有所不同。正确答案总是‘不’]

这些问题是在讲述故事的同时提出的。根据最初的编码标准，参与者必须正确回答所有四个问题，他们的答案才被认为是正确的。然而，在当前的研究中，我们主要感兴趣的是对最后一个问题的回答，测试回答者是否理解说话者的心理状态。在检查人类数据时，我们注意到一些参与者对第一项的反应不正确，因为他们显然不愿意归咎于责任（例如“不，他没有说错任何话，因为他忘记了”）。重点关注与当前研究相关的失礼理解的关键方面

我们将编码限制为仅最后一个问题（1（如果答案是否定的，则正确）或 0（对于其他问题）；请参阅补充信息第 5 节，了解遵循原始标准的替代编码，以及我们编码的重新编码作为正确答案，其中提到正确答案作为可能的解释，但未明确认可）。

以及 Baron-Cohen 等人使用的 10 件原创物品。，我们为本次测试生成了 5 个新颖的项目，它们遵循与原始项目相同的结构和逻辑，总共产生了 15 个项目。

提示任务。暗示任务通过呈现十个按顺序呈现的描述日常社交互动的小插图来评估对间接语音请求的理解。每个小插图都以可解释为提示的评论结尾。

正确的回应既能识别该言论的预期含义，也能识别该言论试图引发的行动。在最初的测试中，如果参与者第一次未能完全回答问题，则会提示他们提出额外的问题。在我们改编的实现中，我们使用 Gil 等人中列出的评估标准删除了这个额外的问题并将答案编码为二进制（1（正确）或 0（不正确））。请注意，此编码提供了比提示理解更保守的估计在之前的研究中。

除了来自 Corcoran 的 10 个原始项目之外，我们还生成了另外 6 个新颖的提示测试项目，总共 16 个项目。

奇怪的故事。这些奇怪的故事提供了一种测试更高级心理能力的方法，例如对误导、操纵、撒谎和误解的推理，以及二阶或更高阶的心理状态（例如，A 知道 B 相信 X……）。这些故事所衡量的高级能力使其适合测试功能较高的儿童和成人。在这个测试中，参与者会看到一个简短的小插曲，并被要求解释为什么一个角色会说或做一些不真实的事情。

每个问题都带有一组特定的编码标准，并且根据其对话语的解释程度以及是否以心智术语对其进行解释，答案可以被授予 0、1 或 2 分。有关部分成功频率的描述，请参阅补充信息第 6 节。

除了 8 个原始心理故事之外，我们还生成了 4 个新颖的项目，总共 12 个项目。可能的最大分数为 24，并且将各个会话分数转换为比例分数进行分析。

测试协议。对于心理理论电池，每次测试都设定了项目的顺序，首先交付原始项目，最后交付新颖项目。每个项目之前都有一个序言，该序言在所有测试中保持一致。接下来是故事描述和相关问题。每个项目交付后，模型都会做出响应，然后会话前进到下一个项目。

对于 GPT 模型，项目是使用聊天 Web 界面交付的。

对于 LLaMA2-Chat 模型，项目的交付是通过自定义脚本自动完成的。对于人类来说，项目在调查的不同页面上显示有自由文本回答框，以便参与者可以写出他们对每个问题的回答（最少字符数为 2）。

失礼可能性检验

为了测试为什么测试模型在失礼测试中表现不佳的替代假设，我们进行了一项仅复制失礼测试的后续研究。该复制遵循与主要研究相同的程序，但有一个主要区别。

该问题的原始措辞是一个简单的是/否问题，测试受试者对说话者错误信念的认识（例如，“理查德记得詹姆斯送给他玩具飞机作为生日礼物吗？”）。为了测试这个问题的低分是否是由于模型在面对歧义时拒绝做出单一解释，我们将其改写为以下术语：

可能性：“理查德更有可能记得还是不记得詹姆斯送给他的生日玩具飞机？”

与原始研究的另一个区别是，在模型未能对错误响应提供清晰推理的极少数情况下，我们添加了后续提示。本次随访的编码标准与其他带有提示系统的研究中使用的编码方案一致，其中无提示的正确答案给出 2 分，提示后的正确答案给出 1 分，提示后的错误答案给出 1 分。给0分。然后将这些分数重新调整为比例分数，以便与原始措辞进行比较。

在人类实验者进行编码的过程中，出现了对不同反应亚型（超出 0-1-2 分）的定性描述，特别是注意到被标记为成功的反应中重复出现的模式。补充信息第 7 节中报告了这种探索性定性细分以及有关提示协议的进一步详细信息。

置信似然检验

为了操纵说话者知道或不知道的可能性，我们开发了一组新的失礼可能性测试变体。对于本对照研究新生成的每个测试项目，我们创建了三个变体：失礼变体、中性变体和知识隐含变体。在失礼变体中，这句话表明说话者不知道上下文。在中性变体中，这句话既不表明他们知道也不表明他们不知道。在隐含知识变体中，话语表明说话者知道（有关所有项目的全文，请参阅补充附录 2）。对于每个变体，核心故事保持不变，例如：

迈克尔在高中时是一个非常笨拙的孩子。 他很难交到朋友，并把时间花在独自写诗上。然而，离开后，他变得更加自信和善于交际。 在他的十年高中同学聚会上，他遇到了阿曼达，她一直在他的英语课上。喝完酒后，她对他说：

接下来是不同情况下的话语：失礼行为：

“我不知道你是否还记得学校里的这个人。他在我的英语课上。他写诗，但他非常尴尬。我希望他今晚不在这里。

中性的：

'Do you know 酒吧在哪儿?

知识蕴涵：

'Do you 仍然 写诗?

置信可能性测试的实施方式与之前的测试相同，不同之处在于反应保持独立，因此不存在反应受到其他变体影响的风险。对于 ChatGPT 模型，这涉及在单独的聊天会话中交付每个项目，每个项目重复 15 次。对于 LLaMA2-70B，这涉及删除 Langchain 对话链，以允许会话内内存上下文。分别招募人类参与者来回答单个测试项目，每个项目至少收集 50 个回答（总数 N = 900）。协议的所有其他细节都是相同的。

量化和统计分析

响应编码。每节课后进行心理理论电池和失礼可能性测试

五位人类实验者根据每个测试预先定义的编码标准对响应进行整理和编码。每个实验者负责为一项测试编写 100% 的会话，为另一项测试编写 20% 的会话。编码器间的一致性百分比是根据 20% 的共享会话计算的，编码器表现出分歧的项目由所有评估者进行评估并重新编码。 OSF 上的可用数据就是此重新编码的结果。如果出现不清楚或不寻常的情况，实验者还会标记个人反应以进行小组评估。评估者间一致性是通过将编码员之间的逐项一致性计算为 1 或 0 并使用它来计算百分比分数来计算的。所有双编码项目的初步一致性超过 95%。最低的一致性是人类和 GPT-3.5 对奇怪故事的反应，但即使在这里，一致性也超过 88%。实验组的委员会评估解决了所有剩余的歧义。

对于信念可能性测试，响应是根据他们是否认可“知道”解释或“不知道”解释，或者他们是否不认可其中一个比另一个更有可能来编码。结果“知道”、“不确定”和“不知道”分别被分配+1、0和-1的数字编码。 GPT 模型在回答时紧密遵循问题的框架，但人类的变化更大，有时会提供模棱两可的回答（例如“是”、“更有可能”和“不太可能”）或者根本不回答问题（“没关系”和“她不在乎”）。这些回答很少见，仅占回答的 2.5%，如果回答是肯定的（“是”），则编码为认可“知道”解释；如果回答是否定的，则编码为“不知道”解释。

统计分析

将 LLMs 与人类表现进行比较。对个人反应的分数进行缩放和平均，以获得每个测试会话的比例分数，以便创建可以在不同心理理论测试中直接比较的绩效指标。我们的目标是将 LLMs 在不同测试中的表现与人类表现进行比较，以了解这些模型在心理理论测试中相对于人类的表现如何。对于每个测试，我们使用一组 Holm 校正的双向 Wilcoxon 测试将三个 LLMs 中的每一个的性能与人类性能进行比较。 Wilcoxon 检验的效应量通过将检验统计量 Z 除以总样本量的平方根来计算，并且效应量的 95% CI 通过 1,000 次迭代进行自举。所有不显着的结果均使用连续先验分布（柯西先验宽度 r = 0.707）下表示为贝叶斯因子 (BF) 的相应贝叶斯检验进行进一步检查。贝叶斯因子在 JASP 0.18.3 中计算，随机种子值为 1。由于上限性能和模型间缺乏方差，错误信念测试的结果未进行推论统计。

新奇的物品。对于每个公开可用的测试（除反讽之外的所有测试），我们生成了新颖的项目，这些项目遵循与原始文本相同的逻辑，但具有不同的细节和文本，以通过包含在 LLM 的性能与已验证的测试项目进行比较。在 JASP 中对不显着的结果进行相应的贝叶斯检验。在新项目上的性能明显低于原始项目，这表明语言模型的良好性能很可能归因于将这些文本包含在训练集中。请注意，虽然暗示和奇怪故事等更复杂任务的开放式格式使其成为这些测试的令人信服的控制，但对于诸如错误信念和失礼之类的任务来说，它们的强度有限，这些任务使用常规的内部结构来进行启发式或“聪明的汉斯”解决方案是可能的。

置信似然检验。我们计算了每个变体和每个模型的不同响应类型（“不知道”、“不确定”和“知道”）的计数频率。然后，对于每个模型，我们进行了两次卡方检验

我们计算了不同响应类型的计数频率（“不知道”测试，将这些分类响应的分布与失礼变体与中性变体进行比较，并将中性变体与隐含知识进行比较。应用了 Holm 校正八个卡方检验以解释多重比较，并使用 JASP 中的贝叶斯列联表进一步检查非显着性结果。

报告摘要

有关研究设计的更多信息，请参阅本文链接的《自然投资组合报告摘要》。

数据可用性

所有资源均可通过知识共享署名非商业 4.0 国际 (CC-BY-NC) 许可证 (https://osf.io/fwj6v) 存储在开放科学框架 (OSF) 上的存储库中。该存储库包含本研究中报告的所有测试项目、数据和代码。测试项目和数据以 Excel 文件形式提供，其中包括每个测试中交付的每个项目的文本、每个项目的全文响应以及分配给每个响应的代码。该文件可从 https://osf.io/dbn92 获取。本文提供了源数据。

代码可用性

主手稿和补充信息中用于所有分析的代码以 Markdown 文件形式包含在 https://osf.io/fwj6v 中。分析文件使用的数据可在存储库中的“scored_data/”下以多个 CSV 文件形式提供，复制分析所需的所有材料都可以在名为“完整 R 项目”的主存储库中作为单个 .zip 文件下载Code.zip'位于 https://osf.io/j3vhq。

参考

1. Van Ackeren, M. J.、Casasanto, D.、Bekkering, H.、Hagoort, P. 和 Rueschemeyer, S.-A.行动中的语用学：间接请求涉及心理理论区域和皮质运动网络。

   J.科格恩。神经科学。 24, 2237–2247 (2012)。
2. Apperly，I.A. 什么是“心智理论”？ 概念、认知过程和个体差异。 Q.J. 实验。心理。 65, 825–839 (2012)。
3. Premack, D. & Woodruff, G. 黑猩猩有心理理论吗？行为。脑科学。 1, 515–526 (1978)。
4. Apperly, I. A.、Riggs, K. J.、Simpson, A.、Chiavarino, C. 和 Samson, D. 信念推理是自动的吗？心理。科学。 17, 841–844 (2006)。
5. 科瓦奇，A. M.、Téglás, E. 和 Endress, A.D. 社会意识：人类婴儿和成人对他人信念的敏感性。

   科学 330, 1830–1834 (2010)。
6. Apperly, I. A.、Warren, F.、Andrews, B. J.、Grant, J. 和 Todd, S. 心智理论的发展连续性：儿童和成人信念-欲望推理的速度和准确性。儿童开发。 82、1691-1703（2011）。
7. Southgate, V.、Senju, A. 和 Csibra, G. 通过 2 岁儿童错误信念的归因进行行动预期。心理。科学。 18, 587–592 (2007)。
8. Kampis, D.、Kármán, P.、Csibra, G.、Southgate, V. 和 Hernik, M. Southgate、Senju 和 Csibra (2007) 的两个实验室直接复制尝试。 R.苏克。打开科学。 8、210190（2021）。
9. 科瓦奇，A. M.、Téglás、E. 和 Csibra, G. 婴儿能否将未明确的内容纳入归因信念？心灵理论的表征先决条件。认知 213, 104640 (2021)。
10. Baron-Cohen, S.、Wheelwright, S.、Hill, J.、Raste, Y. 和 Plumb, I.“从眼睛中读懂内心”测试修订版：一项针对正常成年人和患有阿斯伯格综合症的成年人的研究或者

    高功能自闭症。 J.儿童心理学。精神病学联合学科。

    42, 241–251 (2001)。
11. Wimmer, H. & Perner, J. 关于信念的信念：错误信念在幼儿对欺骗的理解中的代表和约束功能。认知 13, 103–128 (1983)。
12. Perner, J.、Leekam, S. R. 和 Wimmer, H. 三岁儿童的错误信念困难：概念缺陷的情况。 Br。 J.德夫。

    心理。 5, 125–137 (1987)。
13. Baron-Cohen, S.、O’ Riordan, M.、Stone, V.、Jones, R. 和 Plaisted, K. 正常发育儿童和患有阿斯伯格综合症或高功能自闭症儿童的失礼识别。

    J.自闭症开发者。混乱。 29, 407–418 (1999)。
14. Corcoran, R. 归纳推理和精神分裂症意图的理解。认知。神经精神病学 8, 223–235 (2003)。
15. Happé，F.G.E. 心理理论的高级测试：有能力的自闭症、弱智和正常儿童和成人对故事人物的思想和感情的理解。 J.自闭症开发者。混乱。 24, 129–154 (1994)。
16. White, S.、Hill, E.、Happé, F. 和 Frith, U. 重温奇怪的故事：揭示自闭症的心理障碍。儿童开发。

    80、1097-1117（2009）。
17. Apperly, I. A. & Butterfill, S. A. 人类是否有两个系统来追踪信念和类似信念的状态？心理。修订版 116, 953 (2009)。
18. Wiesmann, C. G.、Friederici, A. D.、Singer, T. 和 Steinbeis, N. 在发育中的大脑中思考他人想法的两个系统。过程。国家科学院。科学。美国 117, 6928–6935 (2020)。
19. 布贝克，S.等人。通用人工智能的火花：GPT-4 的早期实验。预印本位于 https://doi.org/10.48550/arXiv.2303.12712 (2023)。
20. 斯里瓦斯塔瓦，A.等人。超越模仿游戏：量化和推断语言模型的能力。预印本位于 https://doi.org/10.48550/arXiv.2206.04615 (2022)。
21. Dou, Z. 探索 GPT-3 模型通过 Sally-Anne 测试的能力两种语言的初步研究。 OSF 预印本 https://doi.org/10.31219/osf.io/8r3ma (2023)。
22. Kosinski, M。心智理论可能自发地出现在大型语言模型中。预印本位于 https://doi.org/10.48550/arXiv.2302.02083 (2023)。
23. Sap, M.、LeBras, R.、Fried, D. 和 Choi, Y. 神经心理理论？关于大型 LM 中社交智能的限制。在过程中。 2022 年自然语言处理经验方法会议 (EMNLP) 3762–3780（计算语言学协会，2022 年）。
24. Gandhi, K.、Fränken, J.-P.、Gerstenberg, T. 和 Goodman, N.D. 理解语言模型中的社会推理

    语言模型。神经信息处理系统进展卷。 36（麻省理工学院出版社，2023）。
25. Ullman, T. 大型语言模型无法对心理理论任务进行微小的改变。预印本位于 https://doi.org/10.48550/arXiv.2302.08399 (2023)。
26. Marcus, G. 和 Davis, E. 如何不测试 GPT-3。马库斯谈人工智能 https://garymarcus.substack.com/p/how-not-to-test-gpt-3 (2023)。
27. 夏皮拉，N.等人。聪明的汉斯还是心理神经理论？在大型语言模型中对社会推理进行压力测试。预印本位于 https://doi.org/10.48550/arXiv.2305.14763 (2023)。
28. 拉万，I.等人。机器行为。自然 568, 477–486 (2019)。
29. Hagendorff, T. 机器心理学：使用心理学方法研究大型语言模型中的涌现能力和行为。预印本位于 https://doi.org/10.48550/arXiv.2303.13988 (2023)。
30. 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。美国 120，e2218523120 (2023)。
31. Webb, T.、Holyoak, K. J. 和 Lu, H. 大语言模型中的紧急类比推理。纳特。哼。行为。 7, 1526–1541 (2023)。
32. Frank, M. C. 公开访问LLMs可以帮助我们理解人类认知。纳特。哼。行为。 7, 1825–1827 (2023)。
33. Bernstein, D. M.、Thornton, W. L. 和 Sommerville, J. A. 历代心智理论：在持续的错误信念任务中，老年人和中年人比年轻人表现出更多的错误。过期。老化研究。 37, 481–502 (2011)。

34. Au-Yeung，S.K.，Kaakinen，J.K.，Liversedge，S.P. 和 Benson，V. 自闭症谱系障碍中书面讽刺的处理：一项眼球运动研究：自闭症谱系障碍中的反讽处理。自闭症研究中心。 8, 749–760 (2015)。

35. Firestone, C. 人机比较中的性能与能力。过程。国家科学院。科学。美国 117, 26562–26571 (2020)。

36. Shapira, N.、Zwirn, G. 和 Goldberg, Y. 大型语言模型在失礼测试中的表现如何？计算语言学协会的调查结果：ACL 2023 10438–10451（计算语言学协会，2023）

37. Rescher, N. 无偏好的选择。对“布里丹的屁股”问题的历史和逻辑的研究。康德梭哈。 51, 142–175 (1960)。

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

39. Chen, L.、Zaharia, M. 和 Zou, J. ChatGPT 的行为如何随时间变化？预印本位于 https://doi.org/10.48550/arXiv.2307.09009 (2023)。

40. Feldman Hall, O. 和 Shenhav, A. 解决社会世界中的不确定性。纳特。哼。行为。 3, 426–435 (2019)。

41. 詹姆斯，W. 心理学原理卷。 2（亨利霍尔特公司，1890）。

42. Fiske, S. T. 思考是为了做：从银版照片到激光照片的社会认知肖像。 J.个人。苏克。心理。 63、877-889（1992）。

43. Plate, R. C.、Ham, H. 和 Jenkins, A. C. 当社会环境中的不确定性增加探索并减少获得的奖励时。

J.Exp。心理。创 152, 2463–2478 (2023)。

44. Frith, C. D. & Frith, U. 心智化的神经基础。神经元 50, 531–534 (2006)。

45. Koster-Hale, J. & Saxe, R. 心理理论：神经预测问题。神经元 79, 836–848 (2013)。

46. 周，P.等人。大型语言模型距离具有心理理论的智能体还有多远？预印本位于 https://doi.org/10.48550/arXiv.2310.03051 (2023)。

47. 博内丰，J.-F. & Rahwan, I. 机器思维，快与慢。

趋势认知。科学。 24, 1019–1027 (2020)。

48. Hanks, T. D.、Mazurek, M. E.、Kiani, R.、Hopp, E. 和 Shadlen, M. N. 经过的决策时间影响感知决策任务中先验概率的权重。 J.神经科学。 31, 6339–6352 (2011)。

49. Pezzulo, G.、Parr, T.、Cisek, P.、Clark, A. 和 Friston, K. 生成意义：主动推理以及被动人工智能的范围和限制。

趋势认知。科学。 28, 97–112 (2023)。

50. Chemero, A. LLMs 与人类认知不同，因为它们不是具体化的。纳特。哼。行为。 7, 1828–1829 (2023)。

51. Brunet-Gouet, E.、Vidal, N. 和 Roux, P. 人类与人工

合理性。 HAR 2023。计算机科学讲义

（Baratgin, J. 等编辑）卷。 14522, 107–126（施普林格，2024）。

52. 金，H.等人。 FANToM：交互中心理理论压力测试的基准。在过程中。 2023 年自然语言处理经验方法会议 (EMNLP) 14397–14413（计算语言学协会，2023 年）。

53. Yiu, E.、Kosoy, E. 和 Gopnik, A. 传播与真理、模仿与创新：儿童可以做到而大型语言和语言与视觉模型还不能做到的事情。透视。心理。科学。

https://doi.org/10.1177/17456916231201401（2023）。

54. Redcay, E. & Schilbach, L. 使用第二人称神经科学来阐明社交互动的机制。纳特。牧师。

神经科学。 20, 495–505 (2019)。

55. 希尔巴赫，L.等人。迈向第二人称神经科学。行为。

脑科学。 36, 393–414 (2013)。

56. Gil, D.、Fernández-Modamio, M.、Bengochea, R. 和 Arrieta, M.

将心理测试的暗示任务理论改编为西班牙语。

普西奎特牧师。祝你健康。英语。埃德。 5, 79–88 (2012)。

致谢

这项工作得到了欧盟委员会通过 ASTOUND 项目（101071191—HORIZON-EIC-2021PATHFINDERCHALLENGES-01 至 A.R.、G.M.、C.B. 和 S.P.）的支持。J.W.A.S.得到了亚历山大・冯・洪堡基金会为经验丰富的研究人员提供的洪堡研究奖学金的支持。资助者在研究设计、数据收集和分析、出版决定或手稿准备中没有任何作用。

作者贡献

J.W.A.S.、A.R.、G.M.、M.S.A.G.和 C.B. 构思了这项研究。 J.W.A.S.、D.A.、G.B.、O.P. 和 E.S.设计任务并执行实验，包括使用人类和 GPT 模型收集数据、响应编码和数据集管理。 S.G.、K.S.和 G.M.从 LLaMA2-Chat 模型收集数据。 J.W.A.S.根据 C.B.、S.P. 和 M.S.A.G. 的意见进行分析并撰写手稿。所有作者都对手稿的解释和编辑做出了贡献。 C.B. 监督了这项工作。 A.R.、G.M.、S.P. 和 C.B. 获得了资金。D.A.、G.B.、O.P. 和 E.S.对工作做出了同等贡献。

资金

由汉堡-埃彭多夫大学 (UKE) 提供开放获取资金。

利益竞争

作者声明没有竞争利益。

附加信息

补充信息 在线版本包含补充材料，网址为 https://doi.org/10.1038/s41562-024-01882-z。

信件和材料请求应发送给 James W. A. Strachan 或 Cristina Becchio。

同行评审信息《自然人类行为》感谢
匿名审稿人对这项工作的同行评审做出的贡献。同行评审报告可供使用。

nature portfolio

通讯作者：詹姆斯·W·A·斯特拉坎 (James W. A. Strachan)克里斯蒂娜·贝奇奥

作者最后更新时间：2024 年 5 月 4 日

报告摘要

Nature Portfolio 希望提高我们发表的作品的可重复性。该表格提供了报告一致性和透明度的结构。有关自然组合政策的更多信息，请参阅我们的编辑政策和编辑政策清单。

统计数据

对于所有统计分析，请确认图形图例、表格图例、正文或方法部分中存在以下项目。

n/a	确认的
☐	☒ 每个实验组/条件的确切样本量 (n)，以离散数字和测量单位给出
☐	☒ 关于测量是否取自不同样品或是否重复测量同一样品的声明
☐	☒ 使用的统计测试以及它们是单侧还是双向 只有常见的测试才应该仅用名称来描述；在方法部分描述更复杂的技术。
☐	☒ 所有测试协变量的描述
☐	☒ 对任何假设或修正的描述，例如正态性检验和多重比较调整
☐	☒ 统计参数的完整描述，包括集中趋势（例如平均值）或其他基本估计（例如回归系数）和变化（例如标准差）或相关的不确定性估计（例如置信区间）
☐	☒ 对于原假设检验，检验统计量（例如 F、t、r）以及置信区间、效应大小、自由度和 P 值均注明。只要合适，请以精确值形式给出 P 值。
☐	☒ 对于贝叶斯分析，有关先验选择和马尔可夫链蒙特卡罗设置的信息
☒	☐ 对于分层和复杂的设计，确定适当的测试级别和完整的结果报告
☐	☒ 效应大小的估计（例如 Cohen's d、Pearson's r），表明它们是如何计算的

我们的生物学家统计网络集包含有关上述许多观点的文章。

软件和代码

有关计算机代码可用性的政策信息

数据收集	人类行为数据是使用 Prolific 平台通过在线实验收集的，该平台针对 SoSci 平台上托管的一项调查。 GPT 模型的数据是通过 http://chat.openai.com 的聊天网络界面收集的。自定义脚本自动交付 LLaMA2-Chat 模型的问题和收集数据，可从 https://www.llama2.ai/ 获取
数据分析	我们使用 R 进行数据分析和创建图形 R 版本 4.1.2 RStudio 2024.04.0-daily+368 “Chocolate Cosmos” Daily (605bbb38ebb4f8565e361122f6d8be3486d288e9, 2024-02-01) 适用于 Ubuntu Jammy Mozilla/5.0 (X11; Linux x86_64 ） AppleWebKit/537.36 （KHTML，如 Gecko） rstudio/2024.04.0-daily+368 Chrome/120.0.6099.56 Electron/28.0.0 Safari/537.36 用于数据分析的代码可作为独立的 RMarkdown 项目从以下位置获取：https： //osf.io/j3vhq 此代码使用以下 R 包： DescTools_0.99.50 flextable_0.9.4 kableExtra_1.3.4 rstatix_0.7.2owplot_1.1.2 ggdist_3.3.1 ggpubr_0.6.0 ggplot2_3.4.4 purrr_1.0.2 Hmisc_5.1-1

tidyr_1.3.0
dplyr_1.1.4
ggtext_0.1.2 主手稿中报告的空结果进行了后续相应的贝叶斯分析，以计算贝叶斯因子（BF10）。此分析是使用 JASP v0.18.3 软件包（JASP 团队，2024 年）

对于使用对研究至关重要但尚未在已发表文献中描述的自定义算法或软件的手稿，必须向编辑和审稿人提供软件。我们强烈鼓励将代码存放在社区存储库（例如 GitHub）中。请参阅 Nature Portfolio 提交代码和软件指南以获取更多信息。

Data

有关数据可用性的政策信息 所有稿件必须包含数据可用性声明。本声明应在适用的情况下提供以下信息： - 公开可用数据集的登录代码、唯一标识符或 Web 链接
- 对数据可用性的任何限制的描述 - 对于临床数据集或第三方数据，请确保声明遵守我们的政策

当前研究中报告的所有数据都可以在知识共享署名非商业 4.0 国际许可证 (CC-BYNC) 下的 OSF 存储库中找到。可以通过以下 URL 访问存储库：https://osf.io/fwj6v/ 问题项的全文、GPT 模型、LLaMA2 模型和人类参与者的响应全文以及分配给每个响应的分数可以从以下 URL 作为单个文件下载：https://osf.io/dbn92 仅包含分数的数据文件（可用于重新创建分析）存储在 OSF 存储库中的 Scored_data/ 文件夹中

涉及人类参与者、他们的数据或生物材料的研究

有关人类参与者或人类数据研究的政策信息。另请参阅有关性别、性别（身份/表现）以及性取向和种族、民族和种族主义的政策信息。

关于性和性别的报告	没有收集性别和性别的数据。
报道种族、民族或其他社会相关群体	没有收集有关种族和民族的数据。
人口特征	我们招募了年龄在 18 岁至 70 岁之间、没有精神疾病史、也没有阅读障碍史的以英语为母语的人。没有收集更多的人口统计数据。
招聘	参与者是通过 Prolific 在线平台招募的，并按调整后的 12 英镑/小时（2 英镑至 6 英镑之间）获得补偿。据我们所知，没有重大的自我选择偏差来源可能会因这一招募程序而影响研究结果。
道德监督	该研究得到了当地伦理委员会（ASL 3 Genovese）的批准，并按照修订后的赫尔辛基宣言的原则进行。

请注意，手稿中还必须提供有关研究方案批准的完整信息。

特定领域的报告

请选择下面最适合您的研究的一项。如果您不确定，请在做出选择之前阅读相应的部分。

☐ 生命科学 ☒ 行为与社会科学 ☐ 生态、进化和环境科学

有关包含所有部分的文档的参考副本，请参阅 nature.com/documents/nr-reporting-summary-flat.pdf

生命科学研究设计

所有研究都必须披露这些观点，即使披露是否定的。

样本量	描述如何确定样本量，详细说明用于预先确定样本量的任何统计方法，或者如果未执行样本量计算，请描述如何选择样本量并提供为什么这些样本量足够的理由。
数据排除	描述任何数据排除。如果没有数据被排除在分析之外，请说明，或者如果数据被排除，请描述排除及其背后的理由，表明是否预先制定了排除标准。
复制	描述为验证实验结果的可重复性而采取的措施。如果所有复制尝试均成功，请确认这一点，或者如果有任何发现未复制或无法复制，请记下这一点并描述原因。
随机化	描述如何将样本/生物体/参与者分配到实验组中。如果分配不是随机的，请描述如何控制协变量，或者如果这与您的研究无关，请解释原因。

致盲	描述研究者在数据收集和/或分析过程中是否对分组分配不知情。如果无法实施盲法，请描述原因或解释为什么盲法与您的研究不相关。

行为与社会科学研究设计

所有研究都必须披露这些观点，即使披露是否定的。

研究描述	这些数据包括对一组心理理论测试问题的全文回答。手稿中报告的数据是根据已发布的编码标准分配给每个文本响应的定量数字分数，与验证程序的任何偏差都在主手稿的方法中明确突出显示。该设计是三种大型语言模型 (LLMs) 与人类受访者基线样本的样本间比较。
研究样本	LLMs：GPT-4、GPT-3.5、LLaMA2-70B（以及补充信息中报告的其他 LLaMA2 模型）：每次测试 15 次（会话）；人类：每次测试的目标为 50 名独特参与者，总数 N=1907（受试者间）。没有收集额外的人口统计信息，但只招募了 18 岁至 70 岁之间、没有阅读障碍或精神疾病史、以英语为母语的人，以确保他们能够完成任务并阅读故事。我们没有指定特定的人口统计数据或收集此数据，因为主要的兴趣比较是人类与LLM性能的比较，我们没有理由建立关于特定人口统计数据的先验假设。招募不限于任何国家，也不限于反映英国或美国人口普查数据的代表性分布。
抽样策略	通过 Prolific 平台提供便捷示例。参与者的参与费用为 12 英镑/小时（2 英镑至 6 英镑之间，具体取决于测试）。样本量是根据 White 等人的对照成人样本量设置的。 (2009)，招募了 40 名神经正常的成年人来更新和验证奇怪的故事任务（作为电池组中最困难的任务，我们认为最有可能表现出变异性）。为了解决在线数据收集带来的任何数据质量问题，我们将每个测试的目标样本大小四舍五入为 N=50。
数据收集	对于每次测试，我们通过 Prolific 收集了每个 LLM 的 15 个会话和约 50 名人类受试者。 GPT 模型通过 OpenAI ChatGPT Web 界面进行测试，会话涉及在同一聊天窗口中交付单个测试的所有项目。 LLaMA模型使用Langchain进行测试，使用设置参数，提示“你是一个有用的AI助手”，温度为0.7，新代币的最大数量设置为512，重复惩罚为1.1，最高P为0.9。对于人类来说，所有项目都是通过 SoSci 平台构建和托管的在线调查按顺序呈现的。实验者对实验条件并不知情，因为与参与者之间没有交互作用。在失礼可能性测试中，实验者在 GPT 模型给出的错误答案推理不清楚的情况下提供后续提示，决定提供后续提示的标准是先验设定的，然后进行评估由其他实验者检查提示是否有效。
定时	主稿和补充材料中报告的全电池 GPT 数据是在 2023 年 4 月 3 日至 4 月 18 日之间收集的。使用 Faux Pas 测试的改编版本的后续数据是在 2023 年 4 月 28 日至 5 月 4 日之间收集的2023 年 4 月 24 日至 5 月 18 日期间收集了使用随机呈现顺序进行讽刺、奇怪故事和失礼测试的 GPT-3.5 后续数据。2023 年 10 月至 11 月期间测试了三个 LLaMA2-Chat 模型。 GPT 模型的错误信念和失礼测试（信念似然测试）于 2023 年 10 月 25 日至 11 月 3 日期间进行。
数据排除	在初步检查数据后，十三 (13) 名人类受试者被排除在最终分析之外。心理理论电池：两 (2) 名受试者使用 GPT 或其他 LLM 来回答问题，一 (1) 名受试者对每个问题都回答“是”；置信可能性测试：七 (7) 名参与者被认为使用 GPT 或其他 LLM 来生成答案；错误信念扰动：三 (3) 名参与者被认为使用 GPT 或其他 LLM 来生成答案。
不参与	没有参与者退出或拒绝参与。
随机化	参与者没有被分配到实验组，而是自愿完成五项心理理论测试之一。这是一个随机机会样本，参加过一次测试的个人被排除在外，不能再次参加。

生态、进化和环境科学研究设计

所有研究都必须披露这些观点，即使披露是否定的。

研究描述	简要描述该研究。对于定量数据，包括处理因素和相互作用、设计结构（例如阶乘、嵌套、分层）、实验单元和重复的性质和数量。
研究样本	描述研究样本（例如一组标记的家雀、器官管仙人掌国家纪念碑内的所有 Stenocereus thurberi），并提供样本选择的理由。如果相关，请描述生物体分类群、来源、性别、年龄范围和任何操作。说明适用时样本代表的人群。对于涉及现有数据集的研究，请描述数据及其来源。
抽样策略	注意采样程序。描述用于预先确定样本量的统计方法，或者如果未执行样本量计算，则描述如何选择样本量并提供为什么这些样本量足够的理由。
数据收集	描述数据收集程序，包括谁记录数据以及如何记录数据。

时间和空间尺度	指出数据收集的开始和停止日期，注意采样的频率和周期，并提供这些选择的理由。如果收集期之间存在间隙，请说明每个样本队列的日期。指定获取数据的空间比例
数据排除	如果没有数据被排除在分析之外，请说明，或者如果数据被排除，请描述排除及其背后的理由，表明是否预先制定了排除标准。
再现性	描述为验证实验结果的可重复性而采取的措施。对于每个实验，请注意重复实验的任何尝试是否失败或说明重复实验的所有尝试均成功。
随机化	描述如何将样本/生物体/参与者分配到组中。如果分配不是随机的，请描述如何控制协变量。如果这与您的研究无关，请解释原因。
致盲	描述数据采集和分析过程中使用的致盲程度。如果无法实施盲法，请描述原因或解释为什么盲法与您的研究不相关。

该研究是否涉及实地工作? ☐ Yes ☐ No

野外工作、收集和运输

现场条件	描述实地工作的研究条件，提供相关参数（例如温度、降雨量）。
地点	说明采样或实验的位置，提供相关参数（例如纬度和经度、海拔、水深）。
访问和导入/导出	描述您为进入栖息地以及以负责任的方式收集和进出口样品所做的努力，并遵守当地、国家和国际法律，并注明所获得的任何许可证（给出颁发机构的名称、日期问题以及任何识别信息）。
扰乱	描述该研究造成的任何干扰以及如何将其最小化。

特定材料、系统和方法的报告

我们需要作者提供有关许多研究中使用的某些类型的材料、实验系统和方法的信息。在此，请指出列出的每种材料、系统或方法是否与您的研究相关。如果您不确定列表项是否适用于您的研究，请在选择响应之前阅读相应的部分。

材料与实验系统

n/a	参与研究
☒	☐ 抗体
☒	☐ 真核细胞系
☒	☐ 古生物学和考古学
☒	☐ 动物和其他生物
☒	☐ 临床数据
☒	☐ 令人关注的双重用途研究
☒	☐ 植物

方法

n/a	参与研究
☒	☐ ChIP测序
☒	☐ 流式细胞仪
☒	☐ 基于 MRI 的神经影像学

抗体

使用的抗体	描述研究中使用的所有抗体；如果适用，请提供供应商名称、目录号、克隆名称和批号。
验证	描述每种一抗针对物种和应用的验证，注意制造商网站上的任何验证声明、相关引文、在线数据库中的抗体概况或手稿中提供的数据。

真核细胞系

有关细胞系以及研究中的性和性别的政策信息

细胞系来源	说明所使用的每种细胞系的来源以及所有原代细胞系和源自人类参与者或脊椎动物模型的细胞的性别。
验证	描述所使用的每个细胞系的验证程序或声明所使用的细胞系均未经过验证。

支原体污染	确认所有细胞系支原体污染测试均为阴性，或描述支原体污染测试结果，或声明细胞系未经过支原体污染测试。
常见的错误识别线路（请参阅 ICLAC 寄存器）	列出研究中使用的任何常见被错误识别的细胞系，并提供其使用的理由。

古生物学和考古学

标本来源	提供标本的出处信息并描述为该工作获得的许可证（包括颁发机构的名称、颁发日期和任何识别信息）。许可证应涵盖收集和（如适用）出口。
标本沉积	注明标本存放地点，以允许其他研究人员自由获取。
约会方法	如果提供了新日期，请描述它们的获取方式（例如收集、储存、样品预处理和测量）、获取地点（即实验室名称）、校准程序和质量保证协议，或说明未提供新日期。

☐ 勾选此框以确认原始日期和校准日期在论文或补充信息中可用。

道德监督	确定批准或提供研究方案指导的组织，或声明不需要伦理批准或指导，并解释为什么不需要。

请注意，手稿中还必须提供有关研究方案批准的完整信息。

动物和其他研究生物

有关动物研究的政策信息；建议报告动物研究以及研究中的性别和性别的 ARRIVE 指南

实验动物	对于实验动物，报告物种、品系和年龄，或说明该研究不涉及实验动物。
野生动物	提供在野外观察或捕获的动物的详细信息；尽可能报告物种和年龄。描述动物是如何被捕获和运输的，以及圈养动物在研究后发生了什么（如果被杀死，请解释原因并描述方法；如果被释放，请说明何时何地）或声明该研究不涉及野生动物。
报告性行为	说明研究结果是否仅适用于一种性别；描述研究设计中是否考虑了性别以及用于分配性别的方法。酌情提供按性别分类的数据，这些信息已在源数据中收集；在本报告摘要中提供总数。如果尚未收集此信息，请说明。报告进行的基于性别的分析，并证明缺乏基于性别的分析的理由。
现场采集的样本	对于使用现场采集样品的实验室工作，请描述所有相关参数，例如外壳、维护、温度、光周期和实验结束协议，或说明该研究不涉及从现场采集的样品。
道德监督	确定批准或提供研究方案指导的组织，或声明不需要伦理批准或指导，并解释为什么不需要。

请注意，手稿中还必须提供有关研究方案批准的完整信息。

临床数据

有关临床研究的政策信息 所有稿件均应符合 ICMJE 临床研究发表指南，并且所有提交的材料中必须包含完整的 CONSORT 检查表。

临床试验注册	提供来自 ClinicalTrials.gov 或同等机构的试验注册号。
研究方案	请注意哪里可以访问完整的试验方案，或者如果不可用，请解释原因。
数据收集	描述数据收集的设置和地点，注意招募和数据收集的时间段。
结果	描述您如何预先定义主要和次要结果指标以及如何评估这些指标。

令人关注的双重用途研究

有关双重用途研究的政策信息

危害

意外、故意或鲁莽地滥用工作中产生的代理或技术，或应用稿件中提供的信息是否会对以下方面构成威胁：

No	Yes	
☐	☐	公共卫生
☐	☐	国家安全
☐	☐	农作物和/或牲畜
☐	☐	生态系统
☐	☐	任何其他重要区域

值得关注的实验

这项工作是否涉及以下任何值得关注的实验：

No	Yes	
☐	☐	演示如何使疫苗失效
☐	☐	赋予对治疗上有用的抗生素或抗病毒药物的耐药性
☐	☐	增强病原体的毒力或使非病原体具有毒力
☐	☐	增加病原体的传播能力
☐	☐	改变病原体的宿主范围
☐	☐	能够逃避诊断/检测方式
☐	☐	实现生物制剂或毒素的武器化
☐	☐	任何其他可能有害的实验和试剂组合

植物

种子库存

报告所有种子库存或使用的其他植物材料的来源。如果适用，请说明种子库存中心和目录号。如果植物标本是从田间采集的，请描述采集地点、日期和采样程序。

新植物基因型

描述产生所有新植物基因型的方法。这包括通过转基因方法、基因编辑、基于化学/辐射的诱变和杂交产生的那些。对于转基因品系，描述转化方法、分析的独立品系的数量以及进行实验的世代。对于基因编辑的细胞系，描述所使用的编辑器、用于编辑的内源序列、靶向引导RNA序列（如果适用）以及如何应用编辑器。

验证

描述所使用的每种种子库或生成的新基因型的任何验证程序。描述用于评估突变影响的任何实验，以及在适用的情况下如何检查潜在的次要影响（例如第二位点 T-DNA 插入、嵌合现象、脱靶基因编辑）。

ChIP测序

数据沉积

☐ 确认原始数据和最终处理数据均已存储在公共数据库（例如 GEO）中。

☐ 确认您已存储或提供了对所调用峰的图形文件（例如 BED 文件）的访问权限。

数据访问链接
发布前可能会保持私密状态。

对于“初始提交”或“修订版”文档，请提供审阅者访问链接。对于您的“最终提交”文档，请提供指向已存数据的链接。

数据库提交中的文件

提供数据库提交中可用的所有文件的列表。

基因组浏览器会话
（例如加州大学圣迭戈分校）

仅针对“初始提交”和“修订版本”文档提供匿名基因组浏览器会话的链接，以进行同行评审。在“最终提交”文件中写上“不再适用”。

方法论

重复

描述实验重复，指定数量、类型和重复协议。

测序深度

描述每个实验的测序深度，提供读数总数、唯一映射的读数、读数长度以及它们是配对还是单端。

抗体

描述用于 ChIP-seq 实验的抗体；如果适用，请提供供应商名称、目录号、克隆名称和批号。

峰值呼叫参数	指定用于读取映射和峰值调用的命令行程序和参数，包括使用的 ChIP、控制和索引文件。
数据质量	详细描述用于确保数据质量的方法，包括有多少个峰处于 FDR 5% 及以上 5 倍富集。
软件	描述用于收集和分析 ChIP-seq 数据的软件。对于已存入社区存储库的自定义代码，请提供加入详细信息。

流式细胞仪

地块

确认：

- [ ] 轴标签说明使用的标记和荧光染料（例如 CD4-FITC）。
- [ ] 轴刻度清晰可见。仅针对组的左下图包含沿轴的数字（“组”是对相同标记的分析）。
- [ ] 所有图都是带有异常值的等高线图或伪彩色图。
- [ ] 提供了细胞数量或百分比（带有统计数据）的数值。

方法论

样品制备	描述样品制备，详细说明细胞的生物来源以及所使用的任何组织处理步骤。
乐器	识别用于数据收集的仪器，指定品牌和型号。
软件	描述用于收集和分析流式细胞术数据的软件。对于已存入社区存储库的自定义代码，请提供加入详细信息。
细胞群丰度	描述分选后组分中相关细胞群的丰度，提供有关样品纯度及其测定方法的详细信息。
门控策略	描述用于所有相关实验的门控策略，指定起始细胞群的初步 FSC/SSC 门，指示定义“阳性”和“阴性”染色细胞群之间的边界的位置。

- [ ] 勾选此框以确认补充信息中提供了门控策略的示例图。

磁共振成像

实验设计

设计类型	指示任务或休息状态；事件相关或块设计。
设计规格	指定每个会话和/或受试者的组块、试验或实验单元的数量，并指定每个试验或组块的长度（如果试验被阻止）以及试验之间的间隔。
行为表现测量	说明记录的变量的数量和/或类型（例如正确的按钮按下、响应时间）以及使用哪些统计数据来确定受试者按预期执行任务（例如受试者之间的平均值、范围和/或标准差）。

获得

成像类型	指定：功能、结构、扩散、灌注。
场强	指定特斯拉
序列和成像参数	指定脉冲序列类型（梯度回波、自旋回波等）、成像类型（EPI、螺旋等）、视野、矩阵大小、切片厚度、方向和 TE/TR/翻转角度。
收购领域	说明是否使用全脑扫描或定义采集区域，描述如何确定该区域。
弥散磁共振成像	- [ ] Used - [ ] 未使用

预处理

预处理软件	提供有关软件版本和修订号以及特定参数（模型/功能、大脑提取、分割、平滑内核大小等）的详细信息。

正常化	如果数据已标准化/标准化，请描述方法：指定线性或非线性并定义用于转换的图像类型或指示数据未标准化并解释缺乏标准化的理由。
标准化模板	描述用于规范化/转换的模板，指定主题空间或组标准化空间（例如 原始 Talairach、MNI305、ICBM152) OR 表示数据未标准化。
噪声和伪影去除	描述您的伪影和结构化噪声消除程序，指定运动参数、组织信号和生理信号（心率、呼吸）。
体积审查	定义体积审查的软件和/或方法和标准，并说明此类审查的范围。

统计建模与推理

模型类型和设置	指定类型（质量单变量、多变量、RSA、预测等）并描述第一和第二级别模型的基本细节（例如固定、随机或混合效应；漂移或自相关）。
测试效果	根据任务或刺激条件而不是心理概念定义精确的效果，并表明是否使用方差分析或因子设计。

指定分析类型：☐ 全脑 ☐ 基于投资回报率 ☐ Both

用于推理的统计类型 （参见 Eklund 等人，2016 年）	指定体素方式或聚类方式，并报告聚类方法的所有相关参数。
更正	描述校正的类型以及如何获得多重比较（例如 FWE、FDR、排列或蒙特卡罗）。

模型与分析

n/a | 参与研究

- ☐ ☐ 功能性和/或有效的连接性
- ☐ ☐ 图表分析
- ☐ ☐ 多变量建模或预测分析

功能性和/或有效的连接性	报告所使用的依赖性度量和模型详细信息（例如皮尔逊相关、偏相关、互信息）。
图表分析	报告因变量和连通性度量，指定加权图或二值化图、主题或组级别以及使用的全局和/或节点摘要（例如聚类系数、效率等）。
多变量建模和预测分析	指定自变量、特征提取和降维、模型、训练和评估指标。