



测试自然语言模型预测人类语言判断的极限

收稿日期：2022 年 6 月 2 日

接受日期：2023 年 8 月 11 日

在线发布：2023 年 9 月 14 日

 检查更新

塔尔戈兰 ^{1,2,6}✉, 马修·西格曼, 尼古拉斯·克里耶斯科特
克里斯托弗·巴尔达萨诺 ³

^{1,3,4,5}

神经网络语言模型似乎越来越符合人类处理和生成语言的方式，但由于语言的离散性和人类语言感知的复杂性，通过对抗性示例识别其弱点具有挑战性。我们通过使模型相互对立来绕过这些限制。我们生成有争议的句子对，其中两个语言模型对于哪个句子更有可能出现存在分歧。考虑到九种语言模型（包括 n-gram、循环神经网络和 Transformer），我们通过综合优化或从语料库中选择句子创建了数百个有争议的句子对。事实证明，有争议的句子对在揭示模型失败和识别与人类对哪个句子更有可能的判断最一致的模型方面非常有效。测试的与人类最一致的模型是 GPT-2，尽管实验也揭示了其与人类感知一致的重大缺陷。

神经网络语言模型不仅是自然语言处理 (NLP) 的关键工具，而且作为人类语言处理的潜在模型也引起了越来越多的科学兴趣。从循环神经网络 (RNN) 到 Transformer，这些语言模型中的每一个（显式或隐式）都定义了单词字符串的概率分布，预测哪些序列可能出现在自然语言中。阅读时间、功能磁共振成像 (fMRI)、头皮脑电图和颅内皮层电图 (ECoG) 等指标的大量证据表明，人类对语言模型捕获的单词和句子的相对概率很敏感，即使是在语法正确且语义有意义。此外，模型导出的句子概率还可以预测人类分级的可接受性判断。然而这些成功

尚未解决两个令人感兴趣的核心问题：（1）哪个模型最符合人类语言处理以及（2）最符合的模型距离完全捕捉人类判断的目标有多近？评估语言模型的主要方法是使用一组标准化基准，例如通用语言理解评估 (GLUE) 或其后继者 SuperGLUE 中的基准。尽管这些基准有助于评估语言模型对下游 NLP 任务的实用性，但事实证明，这些基准不足以比较此类模型作为人类语言处理的候选解释。这些基准的许多组成部分并不是为了测量人类对齐，而是评估模型语言表示在调整到特定下游任务时的有用性。一些基准测试通过比较语言模型分配给语法和非语法句子的概率（例如 BLiMP）来更直接地挑战语言模型。然而，由于此类基准是由理论语言学考虑驱动的，因此它们可能无法检测到语言模型可能偏离人类语言理解的新颖且意想不到的方式。最后，另一个实际问题是，NLP 研究的快速发展导致这些类型的静态基准快速饱和，从而难以区分模型。

1 哥伦比亚大学祖克曼心脑行为研究所，美国纽约州纽约市。以色列贝尔谢巴内盖夫本古里安大学认知与脑科学系。哥伦比亚大学心理学系，美国纽约州纽约市。哥伦比亚大学神经科学系，美国纽约州纽约市。美国纽约州纽约市哥伦比亚大学电气工程系。这些作者做出了同样的贡献：塔尔·戈兰 (Tal Golan)、马修·西格曼 (Matthew Siegelman)。电子邮件：golan.neuro@bgu.ac.il



针对这些问题提出的一个解决方案是使用动态人机交互基准，人们可以通过一组不断发展的测试来主动对模型进行压力测试。然而，这种方法面临的主要障碍是“寻找有趣的例子正在迅速成为一项不那么简单的任务”。我们建议通过模型驱动的评估来补充人工制定的基准。在模型预测而不是实验者直觉的指导下，我们希望识别出信息特别丰富的测试句子，其中不同的模型会做出不同的预测。这种运行经过数学优化的实验以“将特定模型置于危险之中”的方法属于长期存在的设计优化科学哲学。我们可以在大型自然语言语料库中找到这些关键句子，或者合成新颖的测试句子，以揭示不同模型如何泛化到其训练分布之外。

在本文中，我们提出了一种系统的、模型驱动的方法，用于比较语言模型与人类判断的一致性。我们生成有争议的句子对，这些句子对的设计使得两种语言模型对于哪个句子更有可能出现存在强烈分歧。在每个句子对中，一个模型为第一个句子分配的概率高于第二个句子，而另一个模型则更喜欢第二个句子而不是第一个句子。然后，我们收集人类对每对中哪个句子更有可能解决两个模型之间的争议的判断。

这种方法建立在之前关于视觉分类模型有争议图像的工作的基础上。这项工作依赖于对单个刺激的绝对判断，这适用于分类反应。然而，要求参与者在绝对范围内对每个句子的概率进行评分由于幅度估计任务中常见的试验间上下文效应而变得复杂，这已被证明会影响诸如可接受性之类的判断。我们在这里使用的方法是二元强制选择行为任务，让参与者在每次试验中在两个句子之间进行选择，通过在每次试验中设置明确的本地上下文，最大限度地减少试验间上下文效应的作用。这种方法之前已用于测量句子的可接受性，并且与为单个句子提供可接受性评级的设计相比，提供了更多的统计功效。

我们的实验表明（1）可以通过从语料库中选择句子对或迭代修改自然句子来产生有争议的预测，从而为所有常见类别的语言模型生成有争议的句子对；（2）由此产生的有争议的句子对能够在模型之间进行有效的模型比较，而这些模型在人类一致性方面看似等效；（3）所有当前的 NLP 模型类都错误地为某些非自然句子分配了高概率（可以修改自然句子，使其模型概率不会降低，但人类观察者会认为该句子不自然）。这个模型比较和模型测试框架可以让我们对最符合人类语言感知的模型类别有新的见解，并为未来模型的开发提出方向。

结果

我们收集了 100 名以英语为母语的人在线测试的判断。在每次实验中，参与者被要求判断他们“在世界上更有可能遇到的两个句子中的哪一个，无论是语音还是书面文本”，并以三分制的形式对他们的答案的信心进行评分规模（扩展数据图 1 提供了一个试验示例）。该实验旨在比较九种不同的语言模型（补充部分 1.1）：基于二词和三词序列（2-gram 和 3-gram）的语料库频率的概率模型以及包括 RNN 在内的一系列神经网络模型、一个长短期记忆网络 (LSTM) 和五个变压器模型（BERT、RoBERTa、XLM、ELECTRA 和 GPT-2）。

使用自然争议对进行有效的模型比较

作为基线，我们从 Reddit 评论语料库中随机抽样并配对了八个单词的句子。然而，如图 1a 所示，这些句子未能揭示模型之间有意义的差异。对于每个句子对，所有模型都倾向于选择相同的句子（扩展数据图 2），因此在预测人类偏好评级方面表现相似（补充部分 2.1）。

相反，我们可以使用优化过程（补充部分 1.2）来搜索有争议的句子对，其中一个语言模型仅向句子 1 分配高概率（高于自然句子的中值概率），而第二语言模型则为句子 1 分配高概率仅适用于句子 2（示例如表 1 所示）。测量每个模型在预测人类对句子对的选择方面的准确性（其中它是两个目标模型之一），表明模型与人类对齐方面存在许多显着差异（图 1b），其中 GPT-2 和 RoBERTa 显示了最佳的人类一致性，并且 2 克最差。我们还可以单独比较每个模型对（仅使用针对该模型对的刺激），产生类似的成对优势模式（扩展数据图 3a）。除 GPT-2、RoBERTa 和 ELECTRA 之外的所有模型的性能均显着低于我们的噪声上限下限（通过根据其他参与者的响应预测每个参与者的响应而获得的准确度），这表明这些模型的预测与人类判断之间存在偏差。仅在使用有争议的句子对时才显示。

通过合成句子对更好地解开模型

选择有争议的自然句子对可能比随机采样自然句子对提供更大的功效，但该搜索过程考虑了可能句子对空间的非常有限的部分。相反，我们可以迭代地替换自然句子中的单词，以驱动不同的模型做出相反的预测，形成可能位于任何自然语言语料库之外的合成争议句子，如图 2 所示（参见方法，“生成合成争议句子对”了解完整详情）。表 2 显示了对模型预测误差贡献最大的有争议的合成句对的示例。

我们评估了每个模型在所有有争议的合成句子对中预测人类句子选择的效果，其中该模型是两个目标模型之一（图 3a）。这种对模型与人类对齐的评估导致模型的预测精度之间的差距比使用有争议的自然句子对时获得的差距更大，从而使较弱的模型（RNN、3-gram 和 2-gram）远低于 50% 机会准确度级别。在预测人类对这些试验的反应方面，GPT-2、RoBERTa 和 ELECTRA 被发现比替代模型（BERT、XLM、LSTM、RNN、3-gram 和 2-gram）更加准确（比较模型时结果相似）分别配对；扩展数据图 3b）。除 GPT-2 外，所有模型的噪声上限均显着低于下限，这表明与人类判断不一致。

成对的自然句子和合成句子揭示了盲点

最后，我们考虑了一些试验，其中要求参与者在自然句子和从该自然句子生成的合成句子之一之间进行选择。如果语言模型与人类的判断完全一致，我们期望人类同意该模型并至少与自然句子一样选择合成句子。事实上，人类参与者对自然句子表现出比合成句子更系统的偏好（图 3b），即使合成句子的形成使得更强的模型（即 GPT-2、RoBERTa 或 ELECTRA）更倾向于自然句子。自然句子（扩展数据表 1 提供了示例）。分别评估自然句子偏好

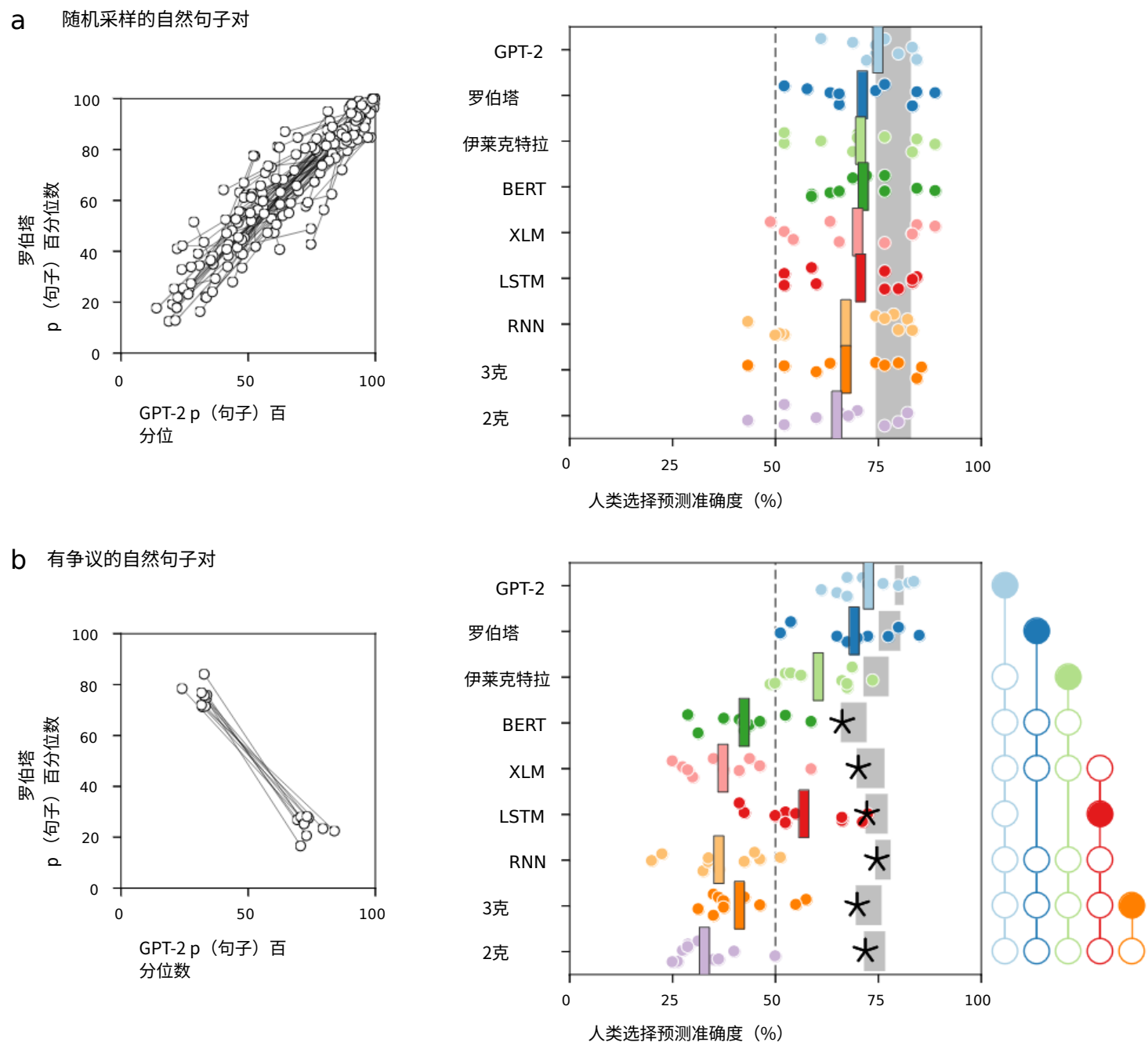


图1|使用自然句子进行模型比较。a，左：百分位数-百分位数散点图，展示了GPT-2和RoBERTa的转换句子概率（相对于实验中使用的所有句子定义）。每对相连的圆圈代表一个句子对。这两个模型在一对内的句子排名方面高度一致（线条具有向上的斜率）。右图：模型预测人类选择的准确性，以模型和人类参与者都偏好同一句子的试验比例来衡量。每个圆圈描绘了一个候选模型的预测准确度，该模型在一组十名参与者中进行了一组独特的试验，并进行了平均。彩色条代表所有100名参与者的总平均值。灰色条是噪声上限，其左右边缘是理想模型将实现的总平均性能的下限和上限（基于人类受试者的一致性）。那里随机采样的自然句子的模型性能没有显著差异。b，左：选择了有争议的自然句子对，使得模型的句子概率等级不一致（线条具有向下的斜率）。右图：有争议的句子对实现了高效的模型比较，揭示了BERT、XLM、LSTM、RNN和n-gram模型的表现显著低于噪声上限（星号表示显著性 - 双边 Wilcoxon 符号秩检验，控制错误发现率在 $q < 0.05$ 处进行九次比较）。在图的右侧，每个实心圆圈表示一个模型显著主导替代模型，用空心圆圈表示（双边 Wilcoxon 符号秩检验，控制所有36个模型对在 $q < 0.05$ 时的错误发现率）。在预测人类判断方面，GPT-2 优于除 RoBERTa 之外的所有模型。

在每个模型配对（扩展数据图 4）中，我们发现即使将强模型与相对弱的模型配对（例如强模型接受合成句子而弱模型拒绝它），这些缺陷也可以被发现。

评估整个数据集揭示了模型的层次结构

我们可以通过计算每个模型相对于我们收集的所有实验的平均预测精度来最大化我们的统计能力，而不是评估每个模型相对于形成的特定句子对的预测精度，以将该模型与替代模型进行比较。此外，这里我们不是将人类和模型的判断二值化，而是测量分级的人类选择（考虑到置信度）与每个候选模型分配的句子概率的对数比之间的序数对应关系。使用这个更敏感的基准（图 4），我们发现 GPT-2 与人类最一致，其次是 RoBERTa，然后是 ELECTRA、BERT、XLM 和 LSTM，以及 RNN、3-gram 和 2-gram 模型。然而

发现所有模型（包括 GPT-2）的准确度明显低于噪声上限下限。与单向变压器 (GPT-2) 相比，双向变压器 (RoBERTa、ELECTRA、BERT 和 XLM) 性能较差的一个可能原因是，在这些模型中计算句子概率很复杂，而且我们开发的概率估计器（参见方法），“评估变压器模型中的句子概率”可能不是最佳的；事实上，流行的伪对数似然 (PLL) 方法对于随机采样的自然句子对的准确度略高（扩展数据图 5a）。然而，当我们通过生成和管理新的合成有争议句子的方式直接将我们的估计器与 PLL 进行比较时，发现我们的估计器明显更符合人类的判断（扩展数据图 5b 和扩展数据表 2）。最后，采用通过标记计数归一化的概率度量的控制分析表明，这种归一化对模型之间观察到的差异的影响很小（补充部分 2.2 和补充图 1）。

表 1 |对每个模型的预测误差影响最大的有争议的自然句子对的示例

句子	对数概率（模型 1）	对数概率（模型 2）	人类选择的数量
n: 生锈一般是由盐和沙子引起的。	$\text{logp}(n \text{GPT-2}) = -50.72$	$\text{logp}(n \text{ELECTRA}) = -38.54$	10
n: 当你需要弗农·罗什时，他在哪里。	$\text{logp}(n \text{GPT-2}) = -32.26$	$\text{logp}(n \text{ELECTRA}) = -58.26$	0
n: 出色的抽吸效果和总体上很棒的吸烟体验。	$\text{logp}(n \text{RoBERTa}) = -67.78$	$\text{logp}(n \text{GPT-2}) = -36.76$	10
n: 我应该更高并与通货膨胀挂钩。	$\text{logp}(n \text{RoBERTa}) = -54.61$	$\text{logp}(n \text{GPT-2}) = -50.31$	0
n: 你可以尝试在他们的论坛上询问。	$\text{logp}(n \text{ELECTRA}) = -51.44$	$\text{logp}(n \text{LSTM}) = -44.24$	10
n: 我喜欢它们看起来像章鱼触手。	$\text{logp}(n \text{ELECTRA}) = -35.51$	$\text{logp}(n \text{LSTM}) = -66.66$	0
n: 成长起来，不要再为一些小不便而抱怨。	$\text{logp}(n \text{BERT}) = -82.74$	$\text{logp}(n \text{GPT-2}) = -35.66$	10
n: 多出来的a是正确的梵文发音。	$\text{logp}(n \text{BERT}) = -51.06$	$\text{logp}(n \text{GPT-2}) = -51.10$	0
n: 出于这个原因，我喜欢我的密码管理器。	$\text{logp}(n \text{XLM}) = -68.93$	$\text{logp}(n \text{RoBERTa}) = -49.61$	10
n: 有点像洞熊氏族。	$\text{logp}(n \text{XLM}) = -44.24$	$\text{logp}(n \text{RoBERTa}) = -67.00$	0
n: 我们培养了一代计算机极客。	$\text{logp}(n \text{LSTM}) = -66.41$	$\text{logp}(n \text{ELECTRA}) = -36.57$	10
n: 我的意思是当裁判粗略时。	$\text{logp}(n \text{LSTM}) = -42.04$	$\text{logp}(n \text{ELECTRA}) = -52.28$	0
n: 这太荒谬了，而且毁了这个爱好。	$\text{logp}(n \text{RNN}) = -100.65$	$\text{logp}(n \text{LSTM}) = -43.50$	10
n: 我觉得男生和无敌更好。	$\text{logp}(n \text{RNN}) = -45.16$	$\text{logp}(n \text{LSTM}) = -59.00$	0
n: 然后用提供的木螺钉将它们固定。	$\text{logp}(n \text{3-gram}) = -119.09$	$\text{logp}(n \text{GPT-2}) = -34.84$	10
n: 听起来你们都被当成狗了。	$\text{logp}(n \text{3-gram}) = -92.07$	$\text{logp}(n \text{GPT-2}) = -52.84$	0
n: 奶油干酪、火腿和洋葱放在饼干上。	$\text{logp}(n \text{2-gram}) = -131.99$	$\text{logp}(n \text{RoBERTa}) = -54.62$	10
n: 我可能必须并行处理饮酒。	$\text{logp}(n \text{2-gram}) = -109.46$	$\text{logp}(n \text{RoBERTa}) = -70.69$	0

对于每个模型（双行，“模型 1”），该表显示了模型严重失败的两个句子的结果。在每种情况下，失败的模型 1 都更喜欢句子 n（粗体对数概率较高），而它所针对的模型（“模型 2”）以及呈现该句子对的所有十个人类受试者都喜欢句子 n。（当超过一对句子在模型中产生相同的最大误差时，表中包含的示例是随机选择的。）

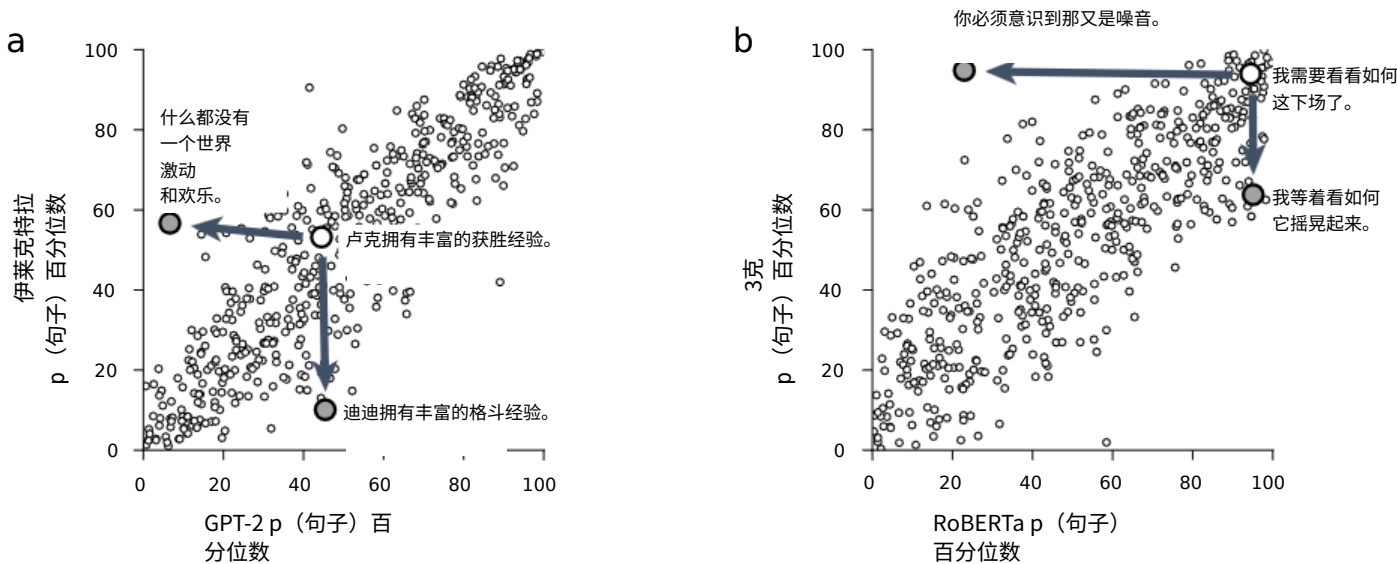


图2|综合有争议的句子对。小空心圆圈表示 500 个随机采样的自然句子。大空心圆圈表示用于初始化有争议的句子优化的自然句子，实心圆圈是最终的合成句子。a，在这个例子中，我们从随机采样的自然句子“Luke has a ton of experience with win”开始。如果我们根据 GPT-2 调整这个句子以最小化其概率（同时根据 ELECTRA 保持该句子至少与自然句子一样可能），我们就得到了合成句子“Nothing has a world of eager and pleasures”。通过在切换角色的同时重复此过程

通过模型，我们生成了合成句子“Diddy has a rich of experience with grappling”，这降低了 ELECTRA 的概率，同时略微增加了 GPT-2 的概率。b，在这个例子中，我们从随机采样的自然句子“我需要看看这是如何进行的”开始。如果我们根据 RoBERTa 调整这个句子以最小化其概率（同时根据 3-gram 保持该句子至少与自然句子一样可能），我们会得到合成句子“You have to recognize is that Noise Again”。如果我们只减少 3-gram 的概率，我们会生成合成句子“I wait to see it how shake out”。

讨论

在这项研究中，我们探讨了语言模型使用有争议的句子对来预测人类相对句子概率判断的能力，这些句子对是经过选择或合成的，因此两个模型对于哪个句子更有可能存在分歧

我们发现 (1) GPT-2（一种用于预测即将到来的标记的单向变压器模型）和 RoBERTa（一种用于保留标记预测任务的双向变压器）对于人类对有争议的自然句子对的判断最具预测性（图1b）； (2) GPT-2、RoBERTa 和 ELECTRA（一种经过训练检测损坏标记的双向变压器）最能预测人类对合成句子对的判断，以最大限度地提高争议性（图 3a）； (3) 考虑到我们收集的整个行为数据集，GPT-2 是最符合人类一致性的模型（图 4）。然而，包括 GPT-2 在内的所有模型都表现出与人类判断不一致的行为——使用替代模型作为反作用力，我们可以

表 2 |对每个模型的预测误差影响最大的有争议的合成句子对的示例

句子	对数概率（模型 1）	对数概率（模型 2）	人类选择的数量
s: 你可以立即了解他的故事。	$\text{logp}(s \text{GPT-2}) = -64.92$	$\text{logp}(s \text{RoBERTa}) = -59.98$	10
s: 任何人都可以斩首拨浪鼓和羚羊。	$\text{logp}(s \text{GPT-2}) = -40.45$	$\text{logp}(s \text{RoBERTa}) = -90.87$	0
s: 但是他们仍然会将你与其他人进行比较。	$\text{logp}(s \text{RoBERTa}) = -53.40$	$\text{logp}(s \text{GPT-2}) = -31.59$	10
s: 为什么人们只把自己奉献给别人。	$\text{logp}(s \text{RoBERTa}) = -48.66$	$\text{logp}(s \text{GPT-2}) = -47.13$	0
s: 他比任何职业运动员康复得都快。	$\text{logp}(s \text{ELECTRA}) = -48.77$	$\text{logp}(s \text{BERT}) = -50.21$	10
s: 一个人的人数少于一支足球队。	$\text{logp}(s \text{ELECTRA}) = -38.25$	$\text{logp}(s \text{BERT}) = -59.09$	0
s: 这就是我们所接受的说法。	$\text{logp}(s \text{BERT}) = -56.14$	$\text{logp}(s \text{GPT-2}) = -26.31$	10
s: 这一周你快要死了。	$\text{logp}(s \text{BERT}) = -50.66$	$\text{logp}(s \text{GPT-2}) = -39.50$	0
s: 早期的逆境使韧性变得更强。	$\text{logp}(s \text{XLM}) = -62.95$	$\text{logp}(s \text{RoBERTa}) = -54.34$	10
s: 一切事物都因无限的尼斯而变得活跃。	$\text{logp}(s \text{XLM}) = -42.95$	$\text{logp}(s \text{RoBERTa}) = -75.72$	0
s: 特朗普总统威胁要袭击白宫。	$\text{logp}(s \text{LSTM}) = -58.78$	$\text{logp}(s \text{RoBERTa}) = -41.67$	10
s: 西萨里拒绝组建白宫。	$\text{logp}(s \text{LSTM}) = -40.35$	$\text{logp}(s \text{RoBERTa}) = -67.32$	0
s: 拉斯豆配上芥末酱味道最好。	$\text{logp}(s \text{RNN}) = -131.62$	$\text{logp}(s \text{RoBERTa}) = -60.58$	10
s: 声明中的大致存活率。	$\text{logp}(s \text{RNN}) = -49.31$	$\text{logp}(s \text{RoBERTa}) = -99.90$	0
s: 你经常看到人们玩多重游戏。	$\text{logp}(s 3\text{-gram}) = -107.16$	$\text{logp}(s \text{ELECTRA}) = -44.79$	10
s: 这大概是最幸福的矛盾伪君子了。	$\text{logp}(s 3\text{-gram}) = -91.59$	$\text{logp}(s \text{ELECTRA}) = -75.83$	0
s: 买家也可以拥有正品。	$\text{logp}(s 2\text{-gram}) = -127.35$	$\text{logp}(s \text{ELECTRA}) = -40.21$	10
s: 我在高中的时候闲逛了一圈。	$\text{logp}(s 2\text{-gram}) = -113.73$	$\text{logp}(s \text{ELECTRA}) = -92.61$	0

对于每个模型（双行，“模型 1”），该表显示了模型严重失败的两个句子的结果。在每种情况下，失败的模型 1 都更喜欢句子 s（粗体对数概率较高），而它所针对的模型（“模型 2”）以及呈现该句子对的所有十个人类受试者都更喜欢句子 s_o。（当超过一对句子在模型中产生相同的最大误差时，表中包含的示例是随机选择的。）

损坏的自然句子，使得它们在模型下的概率不会降低，但人类倾向于拒绝损坏的句子，因为不太可能（图3b）。

对计算心理语言学建模的影响

与卷积神经网络的架构设计原理大致受到生物视觉的启发不同，当前神经网络语言模型的设计在很大程度上不受心理语言学和神经科学的启发。然而，人们正在不断努力采用和调整神经网络语言模型，以作为人类如何处理语言的计算假设，利用各种不同的架构、训练语料库和训练任务。我们发现，一旦与基于 Transformer 的神经网络进行对抗，RNN 就会做出与人类明显不一致的预测。这一发现与最近的证据相一致，即 Transformer 在预测通过 ECoG 或 fMRI 测量的神经反应方面也优于循环网络，并且与基于模型的人类阅读速度和 N400 幅度预测的证据相一致。在变压器中，GPT-2、RoBERTa 和 ELECTRA 表现出了最好的性能。这些模型经过训练仅优化单词级预测任务，与 BERT 和 XLM 不同，BERT 和 XLM 分别针对下一句预测和跨语言任务进行额外训练（并且具有与 RoBERTa 相同的架构）。这表明本地单词预测可以更好地符合人类语言理解。

尽管我们的结果与之前的工作在模型排名方面一致，但 GPT-2 在预测人类对自然与合成争议对的反应方面的重大失败（图 3b）表明 GPT-2 并未完全模拟所使用的计算人类对短句子的处理。这个结果在某种程度上并不令人意外，因为 GPT-2（像我们考虑的所有其他模型一样）是一种现成的机器学习模型，在设计时并未考虑到人类心理语言学和生理细节。然而，我们观察到的人类相当大的不一致似乎与 GPT-2 最近的报告形成鲜明对比，GPT-2 解释了 fMRI 和 ECoG 对自然句子反应的约 100% 的可解释方差。这种差异的部分原因可以解释为 Schrimpf 等人。通过正则化线性回归将 GPT-2 隐藏层激活映射到大脑数据，即使 GPT-2 的整体句子概率与人类不同，它也可以识别 GPT-2 语言表示中的子空间，该子空间与大脑反应非常一致。更重要的是，当用自然语言评估语言模型时，强大的统计模型可能会利用数据中与对人类有意义的特征不同但高度相关的特征。因此，在典型句子上表现良好的模型可能会采用与大脑非常不同的计算机制，这只能通过更具挑战性的领域测试模型来揭示。请注意，即使是我们考虑的最简单的模型 -

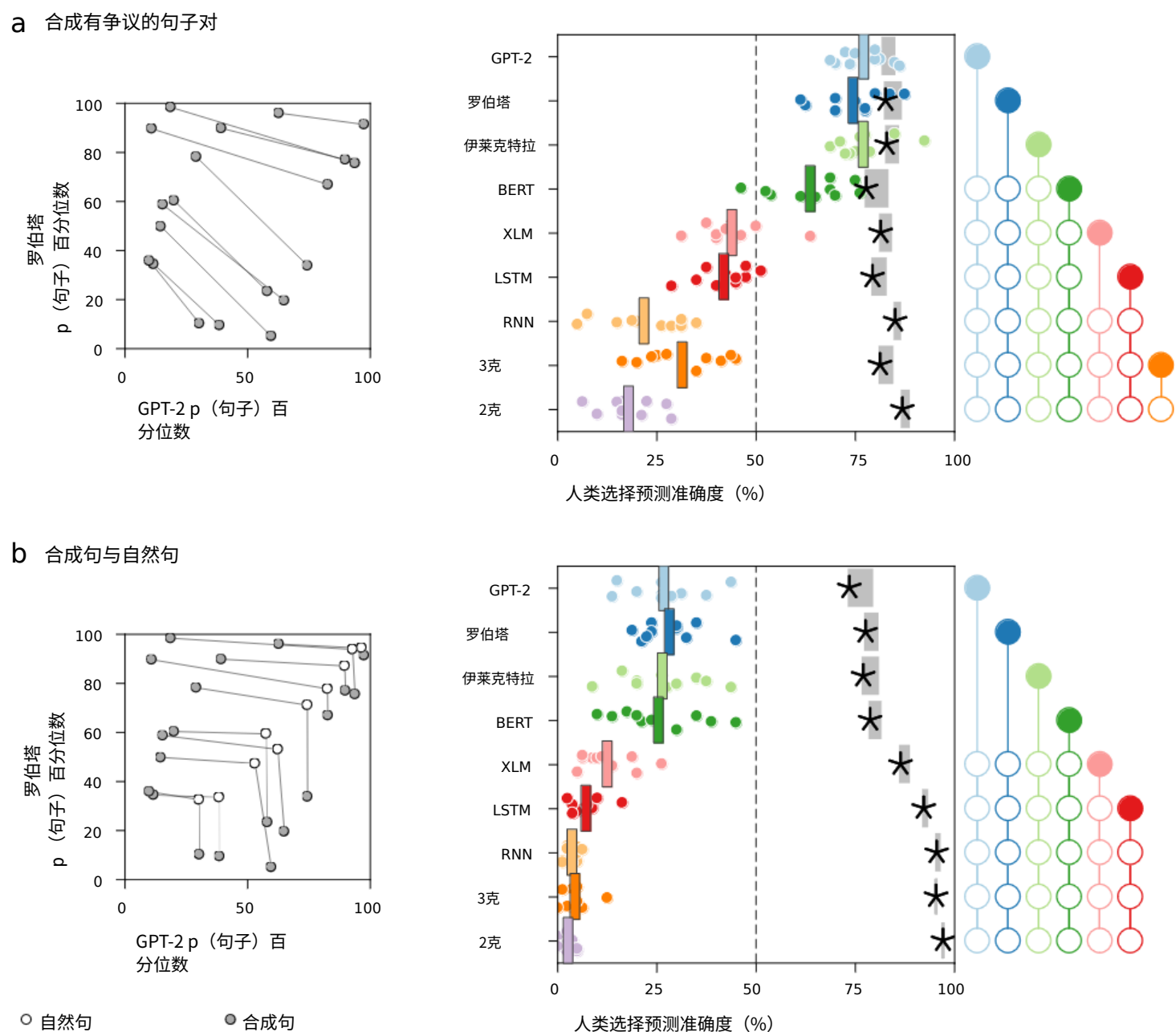


图3|使用合成句子进行模型比较。 a，左：百分位数-对于有争议的合成句子对，GPT-2 和 RoBERTa 的转换句子概率。每对相连的圆圈代表一个句子对。右图：模型预测准确性，以模型和人类参与者都偏好同一句子的试验比例来衡量。GPT-2、RoBERTa 和 ELECTRA 显著优于其他模型（双边 Wilcoxon 符号秩检验，将所有 36 个模型比较的错误发现率控制在 $q < 0.05$ ）。除以下所有型号

三元组圆圈描绘了一个自然句子及其派生的合成句子，经过优化以仅在 GPT-2（三元组中的左侧圆圈）或仅在 RoBERTa（三元组中的底部圆圈）下降低概率。右：每个模型都针对所有合成-自然句子对进行评估，目标是保持合成句子至少与自然句子一样可能（扩展数据图 6 呈现了互补数据分箱）。该评估对所有模型产生了低于机会的预测精度，也显著低于噪声上限的下限。这表明，尽管模型评估这些合成句子至少与原始自然句子一样可能，但人类不同意并表现出对自然句子的系统偏好。有关该图中使用的可视化约定的详细信息，请参见图 1 的标题。

研究发现，GPT-2 的表现低于根据其他参与者的多数票预测每个参与者的选择的噪音上限（灰色）（星号表示显著性 - 双边 Wilcoxon 符号秩检验，控制了 9 个比较的错误发现率） $q < 0.05$ ）。b、左图：各连接

2-gram 频率表——实际上在预测人类对随机采样的自然句子的判断方面表现得相当好，并且只有在受到有争议的句子对的挑战时，它的缺陷才变得明显。我们预测，当使用有意对这种关系进行压力测试的刺激时，使用我们提出的句子级争议性方法或补充想法（例如最大化连续单词之间的争议转移概率），神经表示和当前语言模型之间将会存在很大差异。

使用有争议的句子可以被视为语言模型的泛化测试：模型能否预测自然句子的哪些变化会导致人类拒绝该句子，因为它是不可能的？人类有时能够理解具有非典型结构的语言（例如，在可以根据环境和语言背景对说话者的意图做出语用判断的情况下），但我们测试的模型都不能完全能够预测哪些句法或语义扰动会被人类接受或拒绝。一种可能性是，使用不同的架构、学习规则或训练数据的更强的下一个单词预测模型可能会缩小模型和人类之间的差距。

或者，针对其他语言任务甚至非语言任务需求（特别是代表外部世界、自我和其他主体）的优化可能对于实现类人 NLP 至关重要。

作为对抗性攻击的有争议的句子对

机器视觉模型非常容易受到对抗性例子的影响。此类对抗性示例通常是通过选择正确分类的自然图像，然后搜索会改变目标模型分类的微小（因此人类无法察觉）图像扰动来生成的。语言模型也可能存在类似的隐蔽模型故障模式，这一前景促使人们提出将对抗性方法推广到文本输入。然而，不可察觉的扰动不能应用于书面文本：任何修改的单词或字符都是人类可感知的。先前关于语言模型的对抗性示例的工作依赖于启发式约束，旨在限制文本含义的变化，例如翻转字符、更改数字或性别，或者用同义词替换单词。然而，因为这些启发式只是人类语言的粗略近似

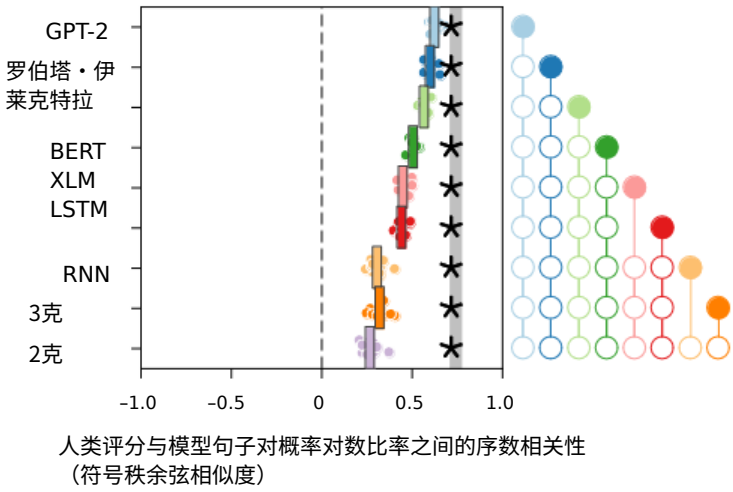


图4|模型句子概率对数比率的序数相关性和人类李克特评级。对于每个句子对，模型预测都被量化

by $\frac{1}{p(s_2|m)}$ 。该对数比率与每个人的李克特评级相关
特定参与者，使用符号秩余弦相似度（方法）。该分析考虑了所有试验和人类置信水平，表明 GPT-2 在预测人类句子概率判断方面表现最佳。

然而，它的预测仍然与人类的选择严重不一致。有关可视化约定的详细信息，请参见图 1 的标题。

处理时，许多这些方法无法保留语义。交互式（“人在环”）对抗方法允许人类受试者反复改变模型输入，从而混淆目标模型，但不会混淆次要参与者，但这些方法本质上是缓慢且昂贵的，并且受到人类受试者心理模型的限制关于评估的语言模型的形式。

相比之下，在有争议的句子对上测试语言模型不需要在优化过程中近似或查询人类的基本事实——争议的目标与正确性无关。相反，通过设计输入来引发模型之间相互冲突的预测，并仅在优化循环终止后评估人类对这些输入的反应，我们利用了一个简单的事实：如果两个模型在输入方面存在分歧，则至少有一个模型一定做出了错误的预测。语言模型与其他语言模型的竞争也可以通过其他方法进行，例如“红队”，其中替代语言模型用作目标模型的潜在对抗性示例的生成器，并且使用分类器来过滤生成的示例例如，它们在目标模型中产生的输出确实是不正确的。我们的方法共享一个基本原则，即替代语言模型可以驱动比手工启发式更强大的测试，但这里的模型具有对称角色（不存在“攻击”和“被攻击”模型），并且我们可以直接优化刺激，而无需依赖过滤。

局限性和未来方向

尽管我们的结果表明，使用有争议的刺激可以识别语言模型与人类判断一致的细微差异，但我们的研究在许多方面受到限制。我们的刺激都是八个单词的英语句子，限制了我们提出适用于全球语言使用的具有认知意义的主张的能力。八个单词的句子足够长，可以包含常见的句法结构并传达有意义的想法，但可能无法有效地探究长距离句法依赖性。未来的工作可能会引入额外的句子长度和语言，以及（可能是自适应的）有争议的句子优化程序，该程序考虑大量候选模型，从而比我们更简单的成对方法允许更大的模型覆盖范围。未来的工作还可以通过旨在识别所有模型共有的潜在故障模式的程序来补充模型比较实验设计。

当前研究的一个更实质性的限制是，就像将预训练的神经网络作为潜在模型进行比较一样

就人类认知而言，特定模型更符合人类判断的原因可能有多种（训练数据、架构、训练任务和学习规则）。例如，由于我们没有系统地控制用于训练模型的训练语料库，因此观察到的一些差异可能是由于训练集而不是模型架构的差异造成的。因此，尽管我们的结果揭示了失败的模型预测，但它们并不能轻易回答为什么会出现这些失败的预测。未来的实验可以比较定制训练或系统操作的模型，这些模型反映了有关人类语言处理的特定假设。在扩展数据图 5 中，我们展示了使用合成的有争议的刺激来对句子概率计算方式存在细微差异的模型之间进行敏感比较的能力。

值得注意的是，我们的分析认为人类相对概率判断反映了可接受性的标量度量。我们做出这个假设是为了将语言模型（为每个句子分配一个概率度量）和人类参与者放在一个共同的基础上。然而，不同类型的句子对可能涉及不同的人类认知过程。对于成对的合成句子，两个句子可能以不同的方式不可接受（例如，表现出不同类型的语法违规），需要权衡多个维度的相对重要性的判断，因此可能会在参与者之间或在试验之间产生不一致的排名。相比之下，要求参与者比较自然句子和合成句子（图 3b 和扩展数据表 1）可能更类似于之前衡量人类对句子对的可接受性判断的工作。尽管如此，值得注意的是，对于所有有争议的条件，噪声上限显着高于模型的预测精度，这表明当前模型无法解释的非随机人类偏好应该由未来模型来解释，这可能必须被考虑在内。更复杂并捕获多个进程。

最后，合成争议句子的使用可以扩展到概率判断之外。足够强大的语言模型可以将实验设计搜索空间限制为特定的句子分布（例如，电影评论或医学问题）。给定这样一个受限的空间，我们也许能够搜索结构良好的句子，这些句子在替代的特定领域模型（例如，情感分类器或问答模型）中引发矛盾的预测。然而，正如我们的结果所表明的，捕获结构良好的句子的分布的任务并不像看起来那么简单。

方法

语言模型

我们测试了来自三个不同类别的九个模型：n-gram 模型、RNN 和 Transformer。n-gram 模型使用 Natural Language Toolkit 中的开源代码进行训练，RNN 使用 PyTorch 中提供的架构和优化程序进行训练，变压器则使用开源存储库 HuggingFace 实现。有关完整详细信息，请参阅补充部分 1.1。

评估 Transformer 模型中的句子概率

然后，我们试图计算上述每个模型下任意句子的概率。“句子”一词在此上下文中以其最广泛的含义使用——一系列英语单词，不一定限于符合语法的英语句子。与某些仅对语法句子（例如情感分析）进行有效模型预测的分类任务不同，句子概率比较任务是在八个单词序列的整个域上定义的。

对于单向模型集，评估句子概率只需将句子中从左到右的每个连续标记的对数概率相加，给定所有

以前的令牌。对于双向模型，这个过程并不那么简单。一个挑战是 Transformer 模型概率不一定反映一致的联合概率；以一种顺序（例如，从左到右）添加单词所产生的对数句子概率之和不一定等于按不同顺序（例如，从右到左）添加单词所产生的概率。在这里，我们开发了一种新颖的双向句子概率公式，其中我们将串行单词位置的所有排列视为可能的构造顺序（类似于用于对串行复制链进行采样的随机单词访问顺序）。在实践中，我们观察到不同排列产生的对数概率分布往往紧密围绕平均值（例如，对于使用自然句子评估的 RoBERTa，平均变异系数约为 0.059）。因此，为了有效计算双向句子概率，我们评估 100 个不同的随机排列，并将整体句子对数概率定义为从每个排列计算的平均对数概率。具体来说，我们初始化了一个八个单词的句子，其中所有标记都替换为在模型训练期间用来代替要预测的单词的“掩码”标记。我们选择位置 1 到 8 的随机排列 P，并在给定其他七个“掩码”标记的情况下，首先计算这些位置 P 中第一个的单词的概率。然后，我们用该位置处的实际单词替换位置 P 处的“掩码”，并计算给定其他六个“掩码”标记和 P 处单词的 P 处单词的概率。重复此过程，直到所有“掩码”标记都已被删除。用相应的词填充。

评估双向 Transformer 模型中句子概率的第二个挑战源于以下事实：这些模型使用单词片段标记器（而不是整个单词），并且这些标记器对于不同的模型是不同的。例如，一个标记器可能将“beehive”一词作为单个标记包含在内，而其他标记器则通过将“beehive”评估为“bee”和“hive”这两个标记来争取较小的独特标记库。多标记词的模型概率（类似于多词句子的概率）可能取决于应用链式法则的顺序。因此，每个多标记词的标记顺序的所有独特排列也在它们各自的掩码内进行评估。例如，单词“beehive”的概率将评估如下：

$$\begin{aligned} &\log p(w = \text{蜂巢}) \\ &= 0.5 (\log p(w = \text{蜜蜂} | w = \text{面具}) + \log p(w = \text{蜂巢} | w = \text{蜜蜂})) \\ &\quad + 0.5 (\log p(w = \text{蜂巢} | w = \text{面具}) + \log p(w = \text{蜜蜂} | w = \text{蜂巢})) \end{aligned} \tag{1}$$

该过程旨在通过（1）确保在周围单词和掩码的相同上下文中评估单词片段标记，从而对单词片段标记的条件概率以及多标记单词的总体概率进行更公平的估计。（2）消除在经过双向预测训练的模型中以任何一种特定顺序评估单词片段标记的偏差。

应用了另一个程序来确保所有模型都计算恰好包含八个单词的句子的概率分布。当使用单词片段标记器评估模型中屏蔽词的条件概率时，我们对模型概率进行归一化，以确保仅考虑单个单词，而不是将屏蔽标记拆分为多个单词。在每个评估步骤中，每个标记都被限制为来自四个标准化分布之一：（1）单掩码单词被限制为带有附加空格的标记，（2）单词开头的掩码被限制为标记带有前置空格（在带有前置空格的模型中，例如 BERT），（3）单词末尾的掩码被限制为带有尾随空格的标记（在带有尾随空格的模型中，例如 XLM）和（4）单词中间的掩码仅限于没有附加空格的标记。

评估令牌计数对句子概率的潜在影响

请注意，由于标记化方案因模型而异，因此句子中的标记数量对于不同模型可能有所不同。这些替代标记化可以被视为建模语言分布的不同分解，改变句子的对数概率在条件概率链上的加性划分方式（但不影响其整体概率）。如果我们尝试通过对数概率除以标记数量来标准化跨模型，就像将模型预测与人类可接受性评级对齐时经常所做的那样，我们的概率将变得强烈依赖于标记化。为了凭经验确认标记化差异不会驱动我们的结果，我们统计比较了每个模型的首选合成句子的标记计数与其非首选对应句子的标记计数。尽管我们发现某些模型存在显着差异，但标记计数和模型句子偏好之间没有系统关联（补充表 1）。特别是，较低的句子概率并没有系统地与较高的标记计数相混淆。

定义共享词汇表

为了促进所有候选模型可以评估的句子的采样、选择和合成，我们定义了包含 29,157 个独特单词的共享词汇表。定义此词汇表对于统一 Transformer 模型（由于其单词片段标记器而可以评估任何输入）与神经网络和 n-gram 模型（包括整个单词作为标记）之间可能的句子空间是必要的，并且确保我们只包含在所有模型的训练语料库中足够流行的单词。词汇表由 SUBTLEX 数据库中的单词组成，删除了用于训练 n-gram 和 RNN 模型（即频率）的 3 亿单词语料库（参见补充部分 1.1）中出现次数少于 300 次的单词。低于百万分之一）。

自然句子的采样

自然句子是从用于构建 n-gram 和 RNN 模型训练语料库的相同四个文本源中采样的，同时确保训练和测试句子之间没有重叠。句子经过过滤，只包含那些具有八个不同单词的句子，除了句子末尾的句号、感叹号或问号之外，没有标点符号。然后，所有八个单词的句子被进一步过滤，以仅包含共享词汇表中包含的单词，并排除包含在预定的不适当单词和短语列表中的单词。为了识别有争议的自然句子对，我们使用整数线性规划来搜索在一个模型中具有高于中值概率且在另一个模型中具有最小概率等级的句子（补充部分 1.2）。

生成合成的有争议的句子对

对于每对模型，我们合成了 100 个句子三元组。每个三元组都用一个自然句子 n（从 Reddit 采样）初始化。根据第一个模型而不是根据第二个模型，对句子 n 中的单词进行迭代修改以生成概率降低的合成句子。重复这个过程，从 n 生成另一个合成句子，其中两个模型的角色颠倒了。从概念上讲，这种方法类似于最大差异化（MAD）竞争，旨在比较图像质量评估模型。每个合成句子都是作为约束最小化问题的解决方案生成的：

$$\begin{aligned} &s = \text{精氨酸} \quad \text{对数 } p(s|m) \\ &\text{服从 } \log p(s|m) \geq \log p(n|m) \end{aligned} \tag{2}$$

其中 m_{reject} 表示目标为与自然句子相比为合成句子分配减少的句子概率的模型， m_{accept} 表示目标为保持合成句子概率大于或等于自然句子的概率的模型。对于一个合成句子，一个模型充当 m_{accept} ，另一个模型充当 m ，而对于另一个合成句子，模型角色被翻转。

在每次优化迭代中，我们伪随机地选择八个单词之一（以便在任何位置被采样 $N + 1$ 次之前，所有八个位置都会被采样 N 次），并在共享词汇表中搜索能够最小化 $\log p(s | m_{reject})$ 在约束下。我们排除了句子中已经出现的潜在替换词，除了允许重复的 42 个限定词和介词（例如“the”、“a”或“with”）。一旦连续八次替换尝试（即未找到减少损失的替换的单词）失败，句子优化过程就结束。

双向模型的字级搜索

对于 $\log p(s|m)$ 评估计算成本较低的模型（2-gram、3-gram、LSTM 和 RNN），我们直接评估 29,157 个可能的单词替换中每一个所产生的 29,157 个句子的对数概率。当这样的概率向量对于两个模型都可用时，我们只需选择最小化损失的替换。对于评估速度较慢的 GPT-2，我们仅评估新单词的条件对数概率（给定句子中的先前单词）不小于 -10 的单词替换的句子概率；在极少数情况下，当该阈值产生的候选单词少于十个时，我们会以五为步减少阈值，直到至少有十个单词高于阈值。对于双向模型（BERT、RoBERTa、XLM 和 ELECTRA），即使对于单个句子， $\log p(s|m)$ 的评估成本也很高，我们使用启发式方法来优先评估要评估的替换。

由于双向模型被训练为掩码语言模型，因此它们很容易提供单词级的完成概率。这些单词级对数概率通常与每个潜在完成所产生的句子的对数概率具有正相关但不完全相关。因此，我们形成了一个基于简单线性回归的 $\log p(s_i \leftarrow w|m)$ 估计，即句子 s 的对数概率，其中单词 w 分配在位置 i ，并根据 $\log p(s_i = w|m)$ 进行预测。 $m, s_i \leftarrow \text{mask}$ ），给定第 i 个单词被屏蔽的句子，单词 w 在位置 i 的完成对数概率：

$$\log \hat{p}(s_i \leftarrow w|m) = \beta \log p(s_i = w|m, s_i \leftarrow \text{掩码}) + \beta$$

(3)

该回归模型是针对每个单词级搜索从头开始估计的。当第一次选择一个单词进行替换时，评估两个句子的对数概率：用完成概率最高的单词替换现有单词得到的句子，以及用完成概率最低的单词替换现有单词得到的句子可能性。这两个单词-句子对数概率对以及当前单词相关的单词-句子对数概率对用于拟合回归线。回归预测与另一个模型的句子概率（如果另一个模型也是双向的，则为精确概率或近似概率）一起用于预测 29,157 个潜在替换中每一个的 $\log p(s|m)$ 。然后，我们用最小预测概率评估替换词的真实（非近似）句子概率。如果这个词确实降低了句子概率，则选择它作为替换词，并终止词级搜索（即继续搜索句子中另一个词的替换词）。如果它没有降低概率，则用新的观察值更新回归模型（方程（3））

并评估期望最小化句子概率的下一个替换。这种词级搜索在五个句子评估后终止，但并没有减少损失。

从优化的句子中选择最好的三元组

因为上述的离散爬山过程是高度局部性的，所以成功地产生高度争议的对的程度根据起始句子 n 的不同而变化。我们发现，通常，具有低于平均对数概率的自然句子会产生具有更大争议的合成句子。为了更好地表示自然句子的分布，同时仍然选择最好的（最具争议的）三元组进行人类测试，我们使用了分层选择。

首先，我们将每个三元组的争议性量化为

$$c(n, s, s) = \text{对数} \frac{p(n|m)}{p(s|m)} + \text{日志} \frac{p(n|m)}{p(s|m)}$$

(4)

其中 s_{is} 是为降低模型 m 中的概率而生成的句子， s_{is} 是为降低模型 m 中的概率而生成的句子。

我们采用整数规划从针对每个模型对优化的 100 个三元组中选择 10 个最具争议性的三元组（最大化所选三元组的总争议性），同时确保对于每个模型，每个十分之一中恰好有一个自然句子自然句子的概率分布。然后使用所选的十个合成三元组来形成每个模型对 30 个独特的实验试验，将自然句子与一个合成句子进行比较，将自然句子与另一个合成句子进行比较，并比较两个合成句子。

人体实验的设计

我们的实验程序得到了哥伦比亚大学机构审查委员会的批准（协议号 IRB-AAAS0252），并按照批准的协议进行。所有参与者均提供了事先知情同意。我们向从 Prolific (www.prolific.co) 招募的 100 名以英语为母语的美国参与者（55 名男性）展示了由语言模型选择和合成的有争议的句子对，并向每位参与者支付 5.95 美元。参与者的平均年龄为 34.08 ± 12.32 岁。受试者被分为十组，每个十组受试者都受到一组独特的刺激。每个刺激集恰好包含模型对的每种可能组合中的一个句子对和四个主要实验条件：选择有争议的句子对；天然与合成对，其中一个模型充当 m_{accept} ，另一个模型充当 m ；自然与合成的配对，具有相反的模型角色分配；并直接将两个合成句子配对。这些模型对条件组合占该任务的 144 (36×4) 次试验。除了这些试验之外，每个刺激集还包括九个试验，这些试验由从八字句子数据库中随机抽样的句子对组成（尚未包含在任何其他条件中）。所有受试者还观看了 12 个对照试验，其中包括随机选择的自然句子和单词以随机顺序打乱的相同自然句子。每个刺激集中的试验顺序以及每个句子对中句子的左右屏幕位置对于所有参与者都是随机的。 尽管优化过程（“生成合成有争议的句子对”部分）产生的每个句子三元组产生了三个试验，但这些试验的分配使得没有受试者两次查看同一个句子。

在每次任务试验中，参与者都被要求对他们认为更有可能的两个句子中的哪一个做出二元决定（有关向参与者提供的全套指令，请参见补充图 2）。此外，他们还被要求表明对自己的决定的三个信心级别之一：有些信心、信心或非常信心。试验没有计时，但整个试验的时间限制为 90 分钟。底部有一个进度条

屏幕向参与者显示他们已经完成了多少次试验以及还需要完成多少次试验。我们拒绝了 21 名参与者的数据，这些参与者在 12 项对照试验中的至少 11 项中未能选择原始的、未打乱的句子，而是从 21 名替代参与者那里获取了数据，所有这些参与者都通过了这一数据质量阈值。总的来说，我们观察到参与者之间的句子偏好高度一致，尽管一致程度因条件而异。对于随机抽样的自然句子对，52.2% 的试验完全或接近完全一致（10 名参与者中至少有 9 人具有相同的二元句子偏好），对于有争议的自然句子对，36.6% 的试验，67.6% 的试验天然-合成对的试验占 60.0%，合成-合成对的试验占 60.0%（假设二项分布 $p = 0.5$ ，机会率为 1.1%）。

模型与人类一致性的评估

为了测量模型判断和人类判断之间每次试验的一致性，我们对这两种测量进行了二值化：我们确定模型为两个句子中的哪一个分配了更高的概率，无论概率差异的大小如何，以及这两个句子中的哪一个无论报告的置信水平如何，句子都受到受试者的青睐。当受试者和模型都选择相同的句子时，该试验被认为是该模型正确预测的。这种正确性测量是对句子对以及观看同一组试验的十名参与者进行平均。对于噪声上限的下限，我们根据接受相同试验的其他九名受试者的多数票来预测每个受试者的选择。对于上限（即该数据样本可达到的最高可能准确度），我们将主题本身包含在这个基于多数投票的预测中。

由于十个参与者组中的每一个都查看了独特的试验集，因此这些组提供了十个独立的实验重复。使用这十个独立的精度结果作为配对样本，通过 Wilcoxon 符号秩检验对模型进行相互比较以及与噪声上限的下限进行比较。对于每项分析，多重比较的错误发现率均由本杰明-霍赫伯格程序控制。

在图 4 中，我们以更连续的方式衡量模型与人类的一致性，将模型中的句子概率比与人类提供的分级李克特评级进行比较（补充部分 1.3 提供了完整的细节）。

选择模型评估的试验

针对每个候选模型对所有随机采样的自然句子对（图 1a）进行了评估。有争议的句子对，无论是自然的（图 1b）还是合成的（图 3），只有在专门针对该模型而形成时才会包含在模型的评估集中。整体总结分析（图 4）评估了所有可用句子对的所有模型。

与伪对数似然可接受性度量的比较

Wang 和 Cho 提出了一种在双向（类似 BERT）模型中计算句子概率的替代方法，使用伪对数似然度量，简单地对以句子中所有其他标记为条件的每个标记的对数概率求和。尽管这一测量并不反映真实的概率分布，但它与人类对几种双向模型的可接受性判断呈正相关。为了直接将这种现有方法与我们计算概率的新方法进行比较，我们再次使用有争议的句子对的方法来识别最符合人类判断的方法。对于每个双向模型（BERT、RoBERTa 和 ELECTRA），我们创建了模型的两个副本

每个都使用不同的方法来计算句子概率。我们合成了 40 个句子对，以最大限度地区分每个模型的两个副本，每个副本为该对中的不同句子分配更高的概率。随后，我们测试了 30 名人类参与者，向每位参与者展示了全部 120 个句子对。与主要实验一样，对模型与人类的一致性进行了量化。

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

实验刺激、详细的行为测试结果以及用于重现所有分析和图表的代码可在 github.com/dpmlab/contstimlang（参考文献 67）上获取。

代码可用性

句子优化代码可在 github.com/dpmlab/contstimlang 上获取（参考文献 67）。

参考

1. Rumelhart, D. E.、Hinton, G. E. 和 Williams, R. J. 通过反向传播误差学习表示。自然 323, 533–536 (1986)。
2. Hochreiter, S. 和 Schmidhuber, J. 长短期记忆。神经计算。9、1735-1780（1997）。
3. Devlin, J.、Chang, M.、Lee, K. 和 Toutanova, K. BERT：用于语言理解的深度双向转换器的预训练。
在过程中。2019 年计算语言学协会北美分会会议：人类语言技术（Burstein, J. 等编辑）4171–4186（计算语言学协会，2019 年）； <https://doi.org/10.18653/v1/n19-1423>
4. 刘, Y.等人。RoBERTa：一种稳健优化的 BERT 预训练方法。预印本 <https://arxiv.org/abs/1907.11692> (2019)。
5. Conneau, A. & Lample, G. 跨语言语言模型预训练。神经信息处理系统的进展（Wallach, H. 等编辑）卷。32（Curran Associates, 2019）； <https://proceedings.neurips.cc/paper/2019/file/c04c19c2c2474dbf5f7ac4372c5b9af1-Paper.pdf>
6. Clark, K.、Luong, M.、Le, Q. V. 和 Manning, C. D. ELECTRA：将预训练文本编码器作为鉴别器而不是发电机。在过程中。第八届国际学习会议 ICLR 2020 陈述（ICLR, 2020）； <https://openreview.net/forum?id=r1xMH1BtvB>
7. 雷德福, A.等人。语言模型是无监督的多任务学习者（OpenAI, 2019）； https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
8. Goodkind, A. 和 Bicknell, K. 单词惊喜对阅读时间的预测能力是语言模型质量的线性函数。
在过程中。第八届认知建模与计算研讨会语言学, CMCL 2018 10–18（计算语言学协会，2018）； <https://doi.org/10.18653/v1/W18-0102>
9. Shain, C.、Blank, I. A.、Schijndel, M.、Schuler, W. 和 Fedorenko, E. fMRI 揭示了自然句子理解过程中特定于语言的预测编码。神经心理学 138, 107307 (2020)。
10. Broderick, M. P.、Anderson, A. J.、Di Liberto, G. M.、Crosse, M. J. 和 Lalor, E. C. 语义差异的电生理相关性反映了对自然叙事语音的理解。电流。生物。28, 803–809 (2018)。
11. 戈尔茨坦, A.等人。人类语言处理和深度语言模型的共享计算原理。纳特。神经科学。25, 369–380 (2022)。

12. Lau, J. H.、Clark, A. 和 Lappin, S. 语法性、可接受性和概率：语言知识的概
率观。认知。
科学。 41, 1202–1241 (2017)。

13. Lau, J. H.、Armendariz, C.、Lappin, S.、Purver, M. 和 Shu, C. 无色的绿色创
意能睡得有多猛烈？句子在上下文中的可接受性。跨。副教授。计算。凌。 8,
296–310 (2020)。

14. 王, A.等人。GLUE：用于自然语言理解的多任务基准测试和分析平台。在过
程中。第七届学习表征国际会议, ICLR 2019 (ICLR, 2019) ； [https://
openreview.net/forum?id=rJ4km2R5t7](https://openreview.net/forum?id=rJ4km2R5t7)

15. 王, A.等人。SuperGLUE：通用语言理解系统的更具粘性的基准。神经信息
处理系统的进展 (Wallach, H. 等编辑) 3266–3280 (Curran Associates,
2019) ； [https://
proceedings.neurips.cc/paper/2019/
file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf](https://proceedings.neurips.cc/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf)

16. 瓦尔施塔特, A.等人。BLiMP：英语语言最小对的基准。跨。副教授。计
算。凌。 8, 377–392 (2020)。

17. 基拉, D.等人。Dynabench：重新思考 NLP 基准测试。在
过程。2021年计算语言学协会北美分会会议：人类语言技术4110–4124
(计算语言学协会, 2021年) ； [https://doi.org/10.18653/v1/2021.
naacl-main.324](https://doi.org/10.18653/v1/2021.naacl-main.324)

18. Box, G. E. P. 和 Hill, W. J. 机械模型之间的区别。技术计量学 9, 57–71
(1967)。

19. Golan, T.、Raju, P. C. 和 Kriegeskorte, N. 有争议的刺激：使神经网络作为
人类认知模型相互对抗。过程。国家科学院。科学。美国 117, 29330–
29337 (2020)。

20. Cross, D. V. 心理物理学判断中的顺序依赖性和回归。感知心理物理学。 14,
547–552 (1973)。

21. Foley, H. J.、Cross, D. V. 和 O’ reilly, J. A. 背景效应的普遍性和幅度：
绝对幅度估计相对性的证据。感知心理物理学。 48, 551–558 (1990)。

22. Petzschner, F. H.、Glasauer, S. 和 Stephan, K. E. 震级估计的贝叶斯视
角。趋势认知。科学。 19, 285–293 (2015)。

23. Greenbaum, S. 对可接受性判断的背景影响。
语言学 15, 5–12 (1977)。

24. Schütze, C. T. 和 Sprouse, J., 《语言学研究方法》(Podesva, R. J. 和
Sharma, D. 编辑) 27–50 (剑桥大学出版社, 2014 年) ； [https://
doi.org/10.1017/CBO9781139013734.004](https://doi.org/10.1017/CBO9781139013734.004)

25. Sprouse, J. & Almeida, D. 可接受性判断实验中的设计灵敏度和统计功效。词
汇 2, 14 (2017)。

26. Lindsay, G. W. 卷积神经网络作为视觉系统的模型：过去、现在和未来。J.
科格恩。神经科学。 33, 2017–2031 (2021) 。

27. Wehbe, L.、Vaswani, A.、Knight, K. 和 Mitchell, T. 将基于上下文的语言统
计模型与大脑活动结合起来
阅读期间。在过程中。2014年自然语言处理经验方法会议 (EMNLP)
233–243 (计算语言学协会, 2014年) ； [https://doi.org/10.3115/v1/
D14-1030](https://doi.org/10.3115/v1/D14-1030)

28. Toneva, M. & Wehbe, L. 用自然语言解释和改进自然语言处理 (在机器
中) -
(在大脑中) 处理。神经信息进展
处理系统 (Wallach, H. 等编辑) 卷。 32 (Curran Associates, 2019) ；
[https://proceedings.neurips.cc/paper/2019/
file/749a8e6c231831ef7756db230b4359c8-Paper.pdf](https://proceedings.neurips.cc/paper/2019/file/749a8e6c231831ef7756db230b4359c8-Paper.pdf)

29. Heilbron, M.、Armeni, K.、Schoffelen, J.-M.、Hagoort, P. 和 De Lange, F.
P. 自然语言理解过程中语言预测的层次结构。过程。国家科学院。科学。美国
119, 2201968119 (2022)。

30. 贾恩, S.等人。可解释的多时间尺度模型，用于预测对连续自然语音的功能磁
共振成像反应。神经信息处理系统的进展 (Larochelle, H. 等编辑)，卷。
33, 13738–13749 (Curran Associates, 2020) ； [https://
process.neurips.cc/paper_files/paper/2020/file/9e9a30b74c
49d07d8150c8c83b1ccf07-Paper.pdf](https://process.neurips.cc/paper_files/paper/2020/file/9e9a30b74c49d07d8150c8c83b1ccf07-Paper.pdf)

31. Lyu, B.、Marslen-Wilson, W. D.、Fang, Y. 和 Tyler, L.K. 寻找时间结构：人
类、机器和语言。预印本位于 [https://www.
biorxiv.org/content/10.1101/2021.10.25.465687v2](https://www.biorxiv.org/content/10.1101/2021.10.25.465687v2) (2021) 。

32. 施林普夫, M.等人。语言的神经架构：集成建模汇聚于预测处理。

过程。国家科学院。科学。美国 118, 2105646118 (2021)。

33. Wilcox, E.、Vani, P. 和 Levy, R. 对神经语言模型和人类增量处理的有针对性的
评估。

在过程中。第59届计算语言学协会年会和第11届自然语言处理国际联合会议
(第一卷：长论文) 939–952 (计算语言学协会, 2021年) ； [https://
doi.org/10.18653/v1/2021.acl-long.76](https://doi.org/10.18653/v1/2021.acl-long.76)

34. Caucheteux, C. 和 King, J.-R. 大脑和算法在自然语言处理中部分融合。交
流。生物。 5, 134 (2022)。

35. Arehalli, S.、Dillon, B. 和 Linzen, T. 神经模型的句法惊喜预测但低估了句法
歧义造成的人类处理难度。在过程中。第 26 届计算自然语言学习会议
(CoNLL) 301–313 (计算语言学协会, 2022 年) ； [https://
aclanthology.org/2022.conll-1.20](https://aclanthology.org/2022.conll-1.20)

36. Merks, D. & Frank, S. L. 人类句子处理：递归
或注意？在过程中。认知建模研讨会
计算语言学 12–22 (计算语言学协会, 2021 年) ； [https://
doi.org/10.18653/v1/2021.cmcl-1.2](https://doi.org/10.18653/v1/2021.cmcl-1.2)

37. Michaelov, J. A.、Bardolph, M. D.、Coulson, S. 和 Bergen, B. K. 不同类型
的认知合理性：为什么 Transformer 在预测 N400 振幅方面比 RNN 更好？在
过程中。认知科学学会年会卷。 43 (2021) ； [https://escholarship.org/
uc/item/9z06m20f](https://escholarship.org/uc/item/9z06m20f)

38. Rakocvic, L. I. 合成有争议的句子来测试语言模型的大脑预测性。博士论文，
麻省理工学院 (2021) ； [https://hdl.handle.
net/1721.1/130713](https://hdl.handle.net/1721.1/130713)

39. Goodman, N. D. & Frank, M. C. 作为概率推理的语用语言解释。趋势认知。科
学。 20, 818–829 (2016)。

40. Howell, S. R.、Jankowicz, D. 和 Becker, S. 扎根语言习得模型：感觉运动特
征改善词汇和语法学习。J.Mem。郎。 53, 258–276 (2005)。

41. 塞格迪, C.等人。神经网络的有趣特性。
预印本位于 <http://arxiv.org/abs/1312.6199> (2013)。

42. Goodfellow, I. J.、Shlens, J. 和 Szegedy, C. 解释和利用对抗性示例。在过程
中。第三届学习表征国际会议, ICLR 2015, 会议记录 (2015) ； [http://
arxiv.org/abs/1412.6572](http://arxiv.org/abs/1412.6572)

43. 张, W.E., 盛Q.Z., Alhazmi, A. & Li, C. 自然语言处理中深度学习模型的对
抗性攻击：a

民意调查。ACM 翻译。英特尔。系统。技术。 11, 1–41 (2020)。

44. 梁, B.等人。深度文本分类可能会被愚弄。在过程中。
第二十七届国际人工智能联合会议
情报, IJCAI-18 4208–4215 (人工智能组织国际联合会议, 2018) ； [https://
doi.org/10.24963/ijcai.2018/585](https://doi.org/10.24963/ijcai.2018/585)

45. Ebrahimi, J.、Rao, A.、Lowd, D. 和 Dou, D. HotFlip：文本分类的白盒对抗示
例。在过程中。计算语言学协会第 56 届年会 (第 2 卷：短论文) 31–36 (计算
语言学协会, 2018 年) ； <https://doi.org/10.18653/v1/P18-2006>

46. 阿卜杜, M.等人。语言模型和人类对 Winograd 模式扰动的敏感性。在过程中。计算语言学协会第58届年会7590–7604（计算语言学协会，2020）；
<https://doi.org/10.18653/v1/2020.acl-main.679>

47. 阿尔赞托特, M.等人。生成自然语言对抗
例子。在过程中。2018年自然语言处理经验方法会议2890–2896（计算语言学协会，2018）；
<https://doi.org/10.18653/v1/D18-1316>

48. Ribeiro, M. T.、Singh, S. 和 Guestrin, C. 用于调试 NLP 模型的语义等效对抗规则。在过程中。
计算语言学协会第 56 届年会（第一卷：长论文）856–865（计算语言学协会，2018 年）；
<https://doi.org/10.18653/v1/P18-1079>

49. Ren, S., Deng, Y., He, K. & Che, W. 通过概率加权词生成自然语言对抗样本

显着性。在过程中。计算语言学协会第57届年会1085–1097（计算语言学协会，2019）；
<https://doi.org/10.18653/v1/P19-1103>

50. Morris, J.、Lifland, E.、Lanchantin, J.、Ji, Y. 和 Qi, Y. 重新评估自然语言中的对抗性示例。计算语言学协会的调查结果：EMNLP 2020 3829–3839（计算语言学协会，2020）；
<https://doi.org/10.18653/v1/2020.findings-emnlp.341>

51. Wallace, E.、Rodriguez, P.、Feng, S.、Yamada, I. 和 Boyd-Graber, J. 如果可以的话，请欺骗我：用于回答问题的人机循环对抗性示例生成。跨。副教授。计算。凌。7, 387–401 (2019)。

52. 佩雷斯, E.等人。红队语言模型与语言
模型。2022 年自然语言处理经验方法会议论文集 3419-3448（计算语言学协会，2022 年）；
<https://doi.org/10.18653/v1/2022.emnlp-main.225>

53. Gibson, E. 语言复杂性：句法依赖的局部性。认知 68, 1–76 (1998)。

54. Watt, W. C. 不可穿透的物体被穿透的不连续性。语言 37, 95–128 (1975)。

55. Schütze, C. T. 语言学的经验基础，经典
语言学卷。2（语言科学出版社，2016）；
<https://doi.org/10.17169/langsci.b89.100>

56. Bird, S.、Klein, E. 和 Loper, E. 使用 Python 进行自然语言处理：使用自然语言工具包分析文本（‘O’ Reilly Media，2009 年）。

57. 帕斯克, A.等人。PyTorch：命令式风格，高性能
深度学习库。神经信息进展
处理系统（Wallach, H. 等编辑）卷。32、8024–8035（Curran Associates, 2019）；
<http://论文.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>

58. 沃尔夫, T.等人。变形金刚：最先进的自然语言
加工。在过程中。2020年自然语言处理经验方法会议：系统演示38-45（计算语言学协会，2020）；
<https://doi.org/10.18653/v1/2020.emnlp-demos.6>

59. Yamakoshi, T.、Griffiths, T. 和 Hawkins, R. 用串行复制链探测 BERT 的先验。计算语言学协会的调查结果，ACL 2022 3977–3992（计算语言学协会，2022）；
<https://doi.org/10.18653/v1/2022.findings-acl.314>

60. Chestnut, S. Perplexity <https://drive.google.com/uc?export=download&id=1gSNfGQ6LPxINctMVwUKrQpUA7OLZ83PW>（2022 年 9 月 23 日访问）。

61. Heuven, W. J. B.、Mandera, P.、Keuleers, E. 和 Brysbaert, M. Subtlex-UK：一个新的、改进的英式英语词频数据库。Q.J. 实验。心理。67、1176-1190（2014）。

62. Wang, Z. & Simoncelli, E. P. 最大微分 (MAD) 竞争：一种比较感知量计算模型的方法。J. 愿景 8, 8 (2008)。

63. Benjamini, Y. & Hochberg, Y. 控制错误发现率：一种实用而强大的多重测试方法。J.R.统计。
苏克。B（方法论）57, 289–300 (1995)。

64. Wang, A. & Cho, K. BERT 有一张嘴，它必须说话：BERT 作为马尔可夫随机场语言模型。在
过程。优化和评估神经语言生成方法研讨会 30-36（计算语言学协会，2019 年）；
<https://doi.org/10.18653/v1/W19-2304>

65. Cho, K. BERT 有一张嘴，必须说话，但它不是 MRF <https://kyunghyuncho.me/bert-has-a-mouth-and-mustspeak-but-it-is-not-an-mrf/>（2022 年 9 月 28 日访问）。

66. Salazar, J.、Liang, D.、Nguyen, T. Q. 和 Kirchhoff, K. Masked
语言模型评分。在过程中。计算语言学协会第58届年会2699-2712（计算语言学协会，2020）；
<https://doi.org/10.18653/v1/2020.acl-main.240>

67. Golan, T.、Siegelman, M.、Kriegeskorte, N. 和 Baldassano, C. “测试预测人类语言判断的自然语言模型的局限性”的代码和数据（Zenodo，2023）；
<https://doi.org/10.5281/zenodo.8147166>

致谢

本材料基于美国国家科学基金会部分资助的工作，资助号为：1948004 转 N.K.本出版物的出版得到了 Charles H. Revson 基金会（T.G.）的支持。然而，所发表的声明和表达的观点仅由作者负责。

作者贡献

T.G.、M.S.、N.K. C.B. 设计了这项研究。多发性硬化症。实现了计算模型和 T.G.实施了句子对优化程序。多发性硬化症。进行了行为实验。T.G.和硕士。分析了实验结果。T.G.、M.S.、N.K. C.B. 撰写了这篇论文。

利益竞争

作者声明没有竞争利益。

附加信息

本文的扩展数据可在 <https://doi.org/10.1038/s42256-023-00718-1> 上获取。

补充信息 在线版本包含补充材料，网址为 <https://doi.org/10.1038/s42256-023-00718-1>。

信件和材料请求应发送给塔尔戈兰。

同行评审信息《自然机器智能》感谢匿名审稿人对这项工作的同行评审做出的贡献。主要处理编辑：Jacob Huth，与自然机器智能团队。

重印和许可信息可在 www.nature.com/reprints 上获取。

自然机器智能 | 第 5 卷 | 2023 年 9 月 | 952–964

963

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。

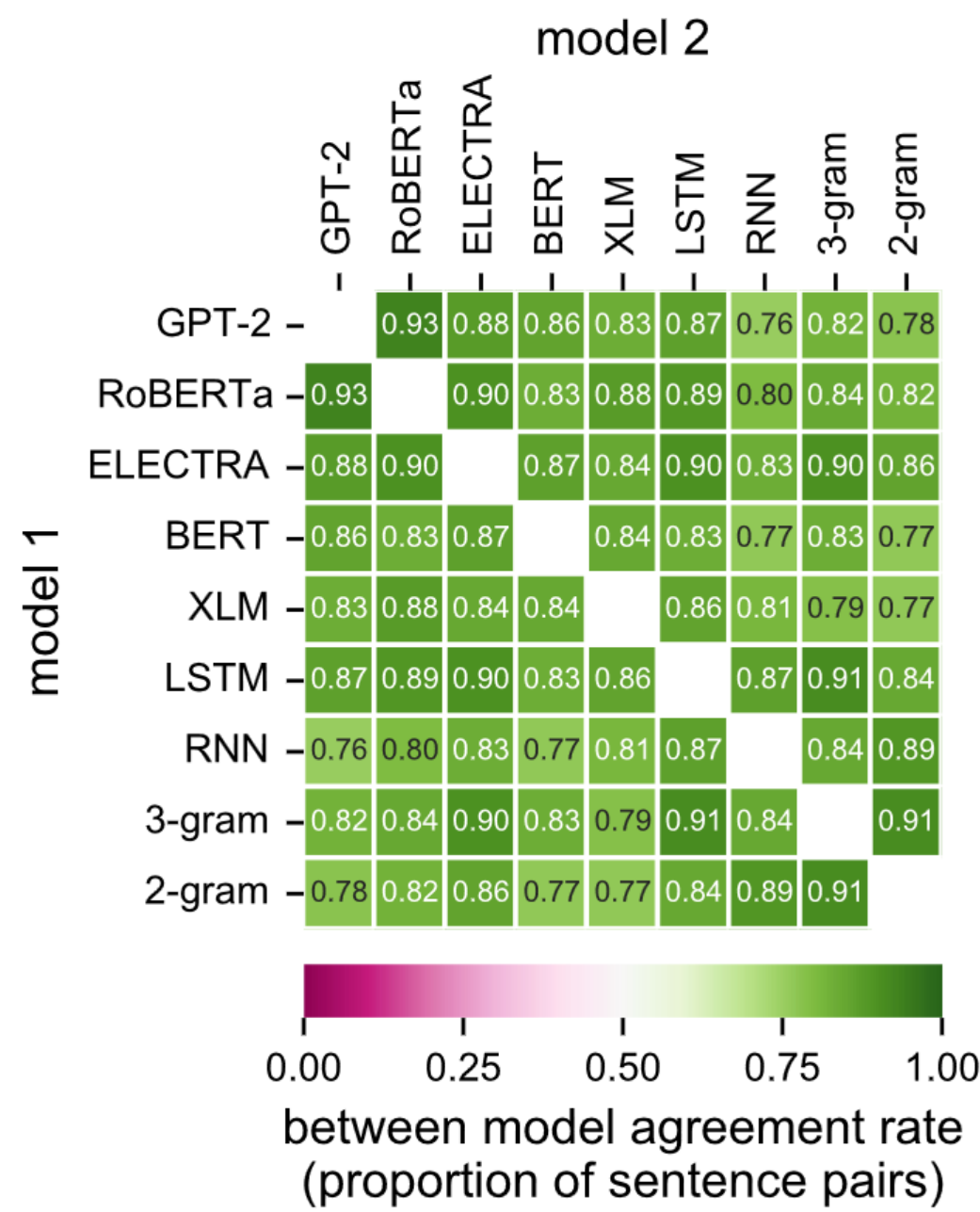
本文已接受的手稿版本仅受此类出版协议条款和适用法律的管辖。

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；作者自存档

© 作者，获得施普林格自然有限公司的独家许可
2023



扩展数据图 1 |向参与者展示的一项实验的示例。参与者必须选择一个句子，同时提供 3 分制的信心评级。

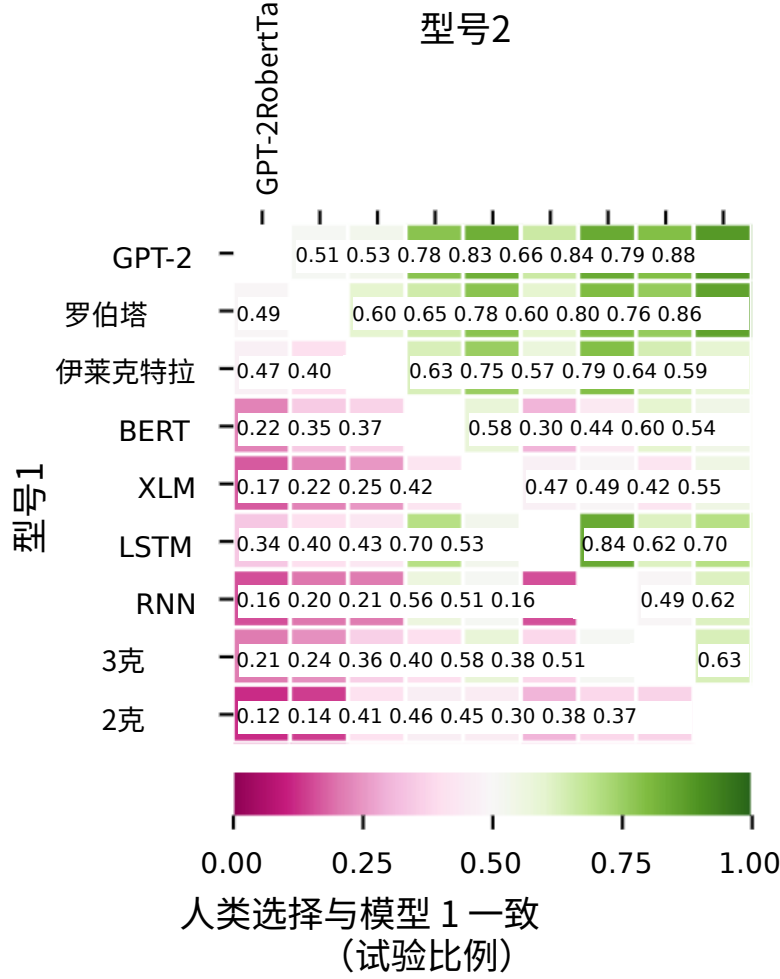


扩展数据图 2 | 90 个随机采样和配对的自然句子对的概率排序的模型间一致性率

两个模型进行一致概率排序的对（即，两个模型都为句子 1 分配较高的概率，或者两个模型都为句子 2 分配较高的概率）。

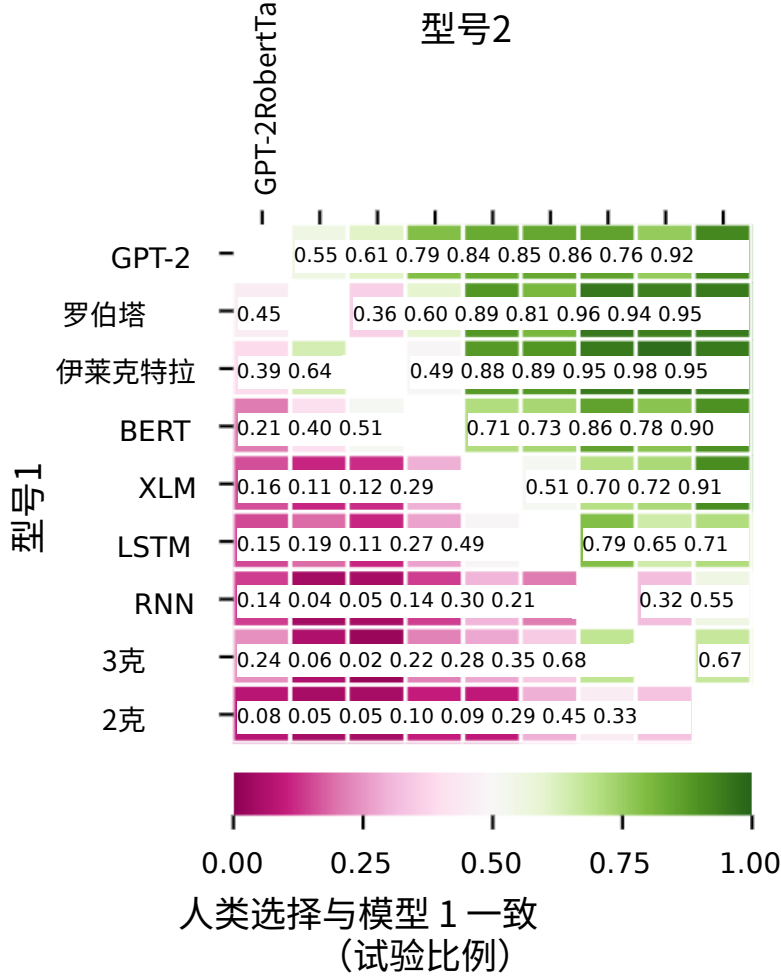
实验中评价。每个单元格代表句子的比例

a 自然争议句

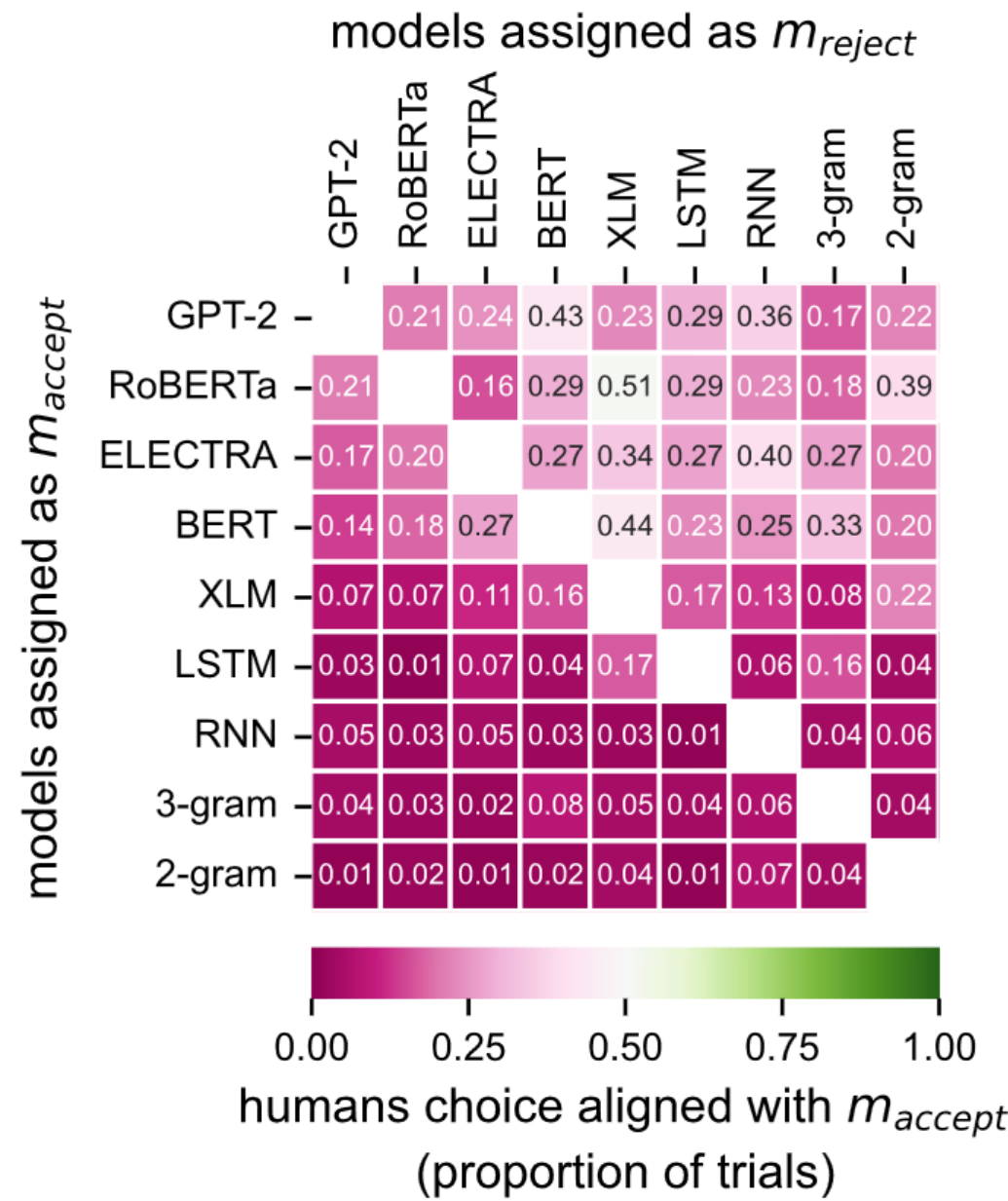


扩展数据图 3 |模型与人类的成对模型比较
一致性。对于每对模型（在上面的矩阵中表示为一个单元格），唯一考虑的试验是选择刺激（a）或合成刺激（b）以对比两个模型的预测的试验。对于这些试验，两个模型总是做出有争议的预测（即第一个模型更喜欢一个句子，而第二个模型更喜欢另一个句子）

b 合成有争议的句子



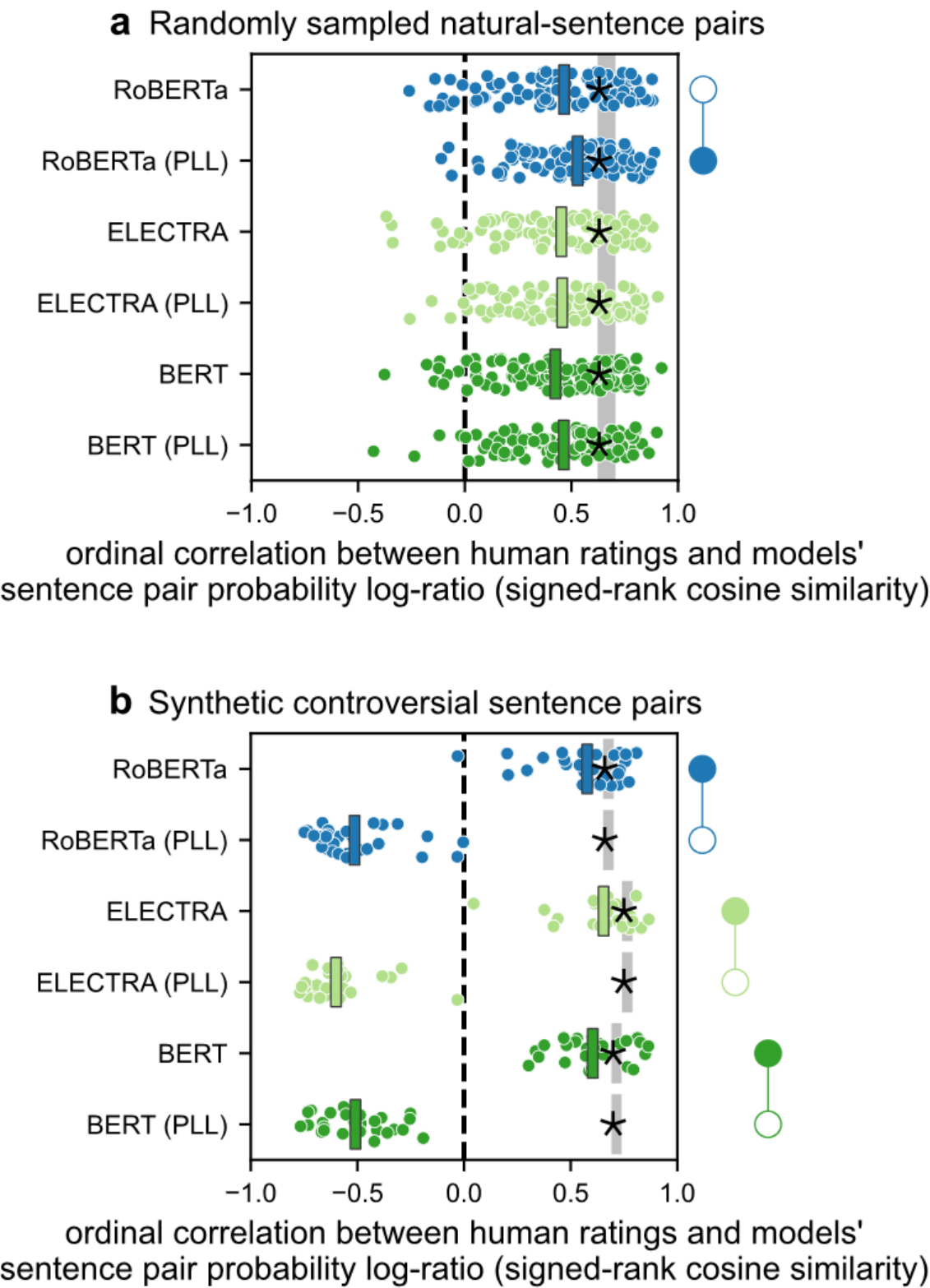
第二个模型）。上面的矩阵描述了二值化人类判断与行模型（“模型 1”）一致的试验比例。例如，GPT-2（顶行）总是比其竞争对手模型更符合人类的选择（绿色色调）。相比之下，与竞争对手模型相比，2-gram（底行）始终不太符合人类的选择（紫色调）。



扩展数据图 4 |人类对自然反应的成对模型分析

与合成句子对。在每个优化条件下，通过修改自然句子 n 形成合成句子 s ，因此合成句子将被一个模型 (m ，列) “拒绝”，最小化 $p(s \mid m)$ ，并被另一个模型 (m ，行)，满足约束 $p(s \mid m) \geq p(n \mid m)$ 。上面的每个单元格总结了由此类优化条件产生的试验中模型与人类的一致性。每个细胞的颜色表示人类判断合成的试验的比例

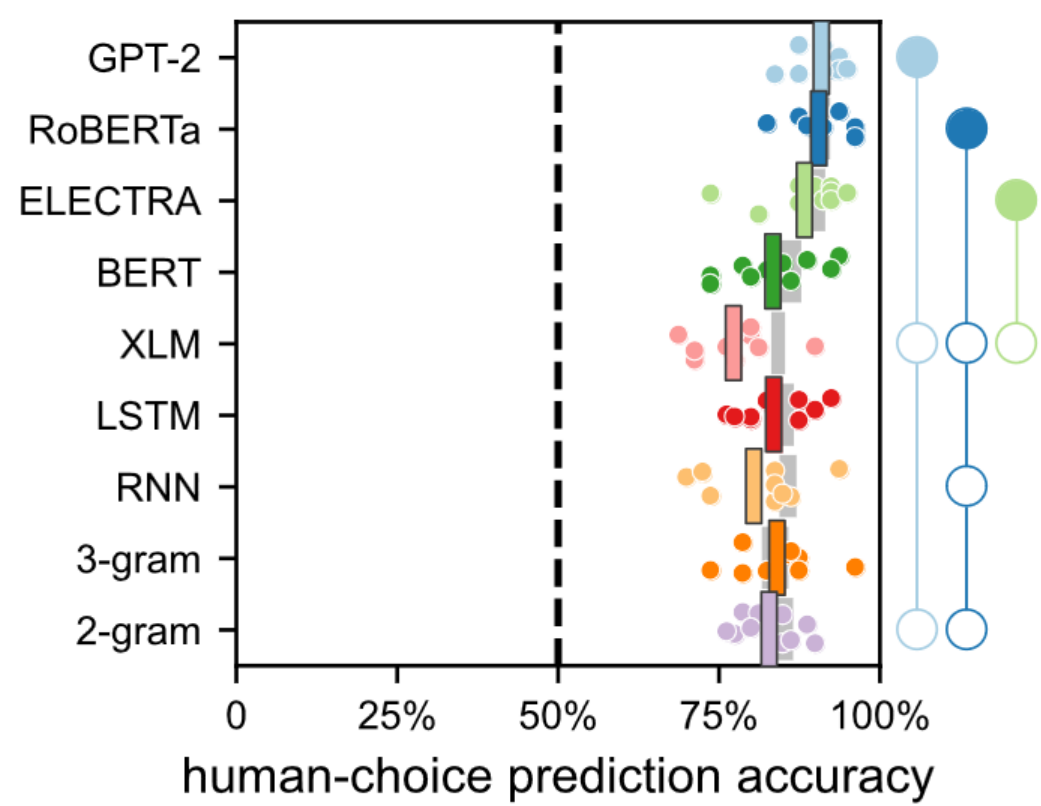
句子比其自然对应句子更有可能，因此与 m 对齐。例如，右上角的单元格描述了人类对句子对的判断，以最小化简单 2-gram 模型分配给合成句子的概率，同时确保 GPT-2 判断合成句子的可能性至少与自然句；在这种情况下，在 100 个句子对中，只有 22 个人类更喜欢合成句子。



扩展数据图 5 |双向变压器的人类一致性：近似对数似然与伪对数似然 (PLL)。上图中的每个点都描述了一名参与者的判断与一个模型的预测之间的序数相关性。(a) BERT、RoBERTa 和 ELECTRA 在主要实验中使用两种不同的似然度量来预测人类对随机采样的自然句子对的判断的性能：我们新颖的近似似然方法（即对多个条件概率链进行平均，参见方法）和伪似然（PLL，在给

定所有其他单词的情况下对每个单词的概率求和）。对于每个模型，我们使用双边 Wilcoxon 符号秩检验对参与者进行统计比较，将两个似然度量相互进行比较，并与噪声上限进行比较。9 次比较的错误发现率控制在 $q < 0.05$ 。在预测人类对自然句子的偏好时，

伪对数似然度量至少与我们提出的近似对数似然度量一样准确。(b) 后续实验的结果，其中我们为每个模型对合成了合成句子对，使两种替代似然性度量相互竞争。统计测试以与面板a中相同的方式进行。这些结果表明，对于三个双向语言模型中的每一个，近似对数似然度量比伪似然度量更加人类一致 ($q < 0.05$)。合成有争议的句子对揭示了伪对数似然度量的戏剧性失败模式，当评估仅限于随机采样的自然句子时，这种模式仍然是隐蔽的。有关合成句对示例，请参阅扩展数据表 2。



扩展数据图 6 |自然句子和合成句子对的模型预测准确性，评估所有句子对中的每个模型，目标是将合成句子的概率低于自然句子。这里应用的数据分箱是对图 3b 中使用的数据分箱的补充，其中每个模型都在所有模型中进行评估。

与自然句子的可能性最小。与图 3b 不同，图 3b 中所有模型都表现不佳，但没有发现模型明显低于噪声上限下限；通常，当一个句子被优化以降低其在任何模型下的概率时（尽管在第二个模型下句子概率没有降低），人们一致认为该句子的概率变得更小。

目标是对合成句子进行评分的句子对

扩展数据表 1 | 对每个模型的预测误差贡献最大的合成句子和自然句子对的示例

sentence	log probability (model 1)	log probability (model 2)	# human choices
<i>n</i> : I always cover for him and make excuses. <i>s</i> : We either wish for it or ourselves do.	$\log p(n \text{GPT-2}) = -36.46$ $\log p(s \text{GPT-2}) = \mathbf{-36.15}$	$\log p(n 2\text{-gram}) = \mathbf{-106.95}$ $\log p(s 2\text{-gram}) = -122.28$	10 0
<i>n</i> : This is why I will never understand boys. <i>s</i> : This is why I will never kiss boys.	$\log p(n \text{RoBERTa}) = -46.88$ $\log p(s \text{RoBERTa}) = \mathbf{-46.75}$	$\log p(n 2\text{-gram}) = \mathbf{-103.11}$ $\log p(s 2\text{-gram}) = -107.91$	10 0
<i>n</i> : One of the ones I did required it. <i>s</i> : Many of the years I did done so.	$\log p(n \text{ELECTRA}) = -35.97$ $\log p(s \text{ELECTRA}) = \mathbf{-35.77}$	$\log p(n \text{LSTM}) = \mathbf{-40.89}$ $\log p(s \text{LSTM}) = -46.25$	10 0
<i>n</i> : There were no guns in the Bronze Age. <i>s</i> : There is rich finds from the Bronze Age.	$\log p(n \text{BERT}) = -48.48$ $\log p(s \text{BERT}) = \mathbf{-48.46}$	$\log p(n \text{ELECTRA}) = \mathbf{-30.40}$ $\log p(s \text{ELECTRA}) = -44.34$	10 0
<i>n</i> : You did a great job on cleaning them. <i>s</i> : She did a great job at do me.	$\log p(n \text{XLM}) = -40.38$ $\log p(s \text{XLM}) = \mathbf{-39.89}$	$\log p(n \text{RNN}) = \mathbf{-43.47}$ $\log p(s \text{RNN}) = -61.03$	10 0
<i>n</i> : This logic has always seemed flawed to me. <i>s</i> : His cell has always seemed instinctively to me.	$\log p(n \text{LSTM}) = -39.77$ $\log p(s \text{LSTM}) = \mathbf{-38.89}$	$\log p(n \text{RNN}) = \mathbf{-45.92}$ $\log p(s \text{RNN}) = -62.81$	10 0
<i>s</i> : Stand near the cafe and sip your coffee. <i>n</i> : Sit at the front and break your neck.	$\log p(s \text{RNN}) = -65.55$ $\log p(n \text{RNN}) = \mathbf{-44.18}$	$\log p(s \text{ELECTRA}) = \mathbf{-34.46}$ $\log p(n \text{ELECTRA}) = -34.65$	10 0
<i>n</i> : Most of my jobs have been like this. <i>s</i> : One of my boyfriend have been like this.	$\log p(n 3\text{-gram}) = -80.72$ $\log p(s 3\text{-gram}) = \mathbf{-80.63}$	$\log p(n \text{LSTM}) = \mathbf{-35.07}$ $\log p(s \text{LSTM}) = -41.44$	10 0
<i>n</i> : They even mentioned that I offer white flowers. <i>s</i> : But even fancied that would logically contradictory philosophies.	$\log p(n 2\text{-gram}) = -113.38$ $\log p(s 2\text{-gram}) = \mathbf{-113.24}$	$\log p(n \text{BERT}) = \mathbf{-62.81}$ $\log p(s \text{BERT}) = -117.98$	10 0

对于每个模型（双行，“模型 1”），该表显示了模型严重失败的两个句子的结果。在每种情况下，失败的模型 1 更喜欢合成句子 s（较高的对数概率以粗体显示），而它所针对的模型（“模型 2”）以及呈现该句子对的所有 10 名人类受试者更喜欢自然句子 n。（当超过一对句子在模型中产生相同的最大误差时，表中包含的示例是随机选择的。

扩展数据表 2 |使用伪对数似然 (PLL) 对双向变压器的预测误差产生最大影响的有争议的合成句子对的示例

sentence	pseudo-log-likelihood (PLL)	approximate log probability	# human choices
s_1 : I found so many in things and called.	$\log p(s_1 \text{BERT (PLL)}) = -55.14$	$\log p(s_1 \text{BERT}) = -\mathbf{55.89}$	30
s_2 : Khrushchev schizophrenic so far disproportionately goldfish fished alone.	$\log p(s_2 \text{BERT (PLL)}) = -\mathbf{22.84}$	$\log p(s_2 \text{BERT}) = -162.31$	0
s_1 : Figures out if you are on the lead.	$\log p(s_1 \text{BERT (PLL)}) = -38.11$	$\log p(s_1 \text{BERT}) = -\mathbf{51.27}$	30
s_2 : Neighbours unsatisfactory indistinguishable misinterpreting schizophrenic on homecoming cheerleading.	$\log p(s_2 \text{BERT (PLL)}) = -\mathbf{16.43}$	$\log p(s_2 \text{BERT}) = -258.91$	0
s_1 : I just say this and not the point.	$\log p(s_1 \text{ELECTRA (PLL)}) = -34.41$	$\log p(s_1 \text{ELECTRA}) = -\mathbf{33.80}$	30
s_2 : Glastonbury reliably mobilize disenfranchised homosexuals underestimate unhealthy skeptics.	$\log p(s_2 \text{ELECTRA (PLL)}) = -\mathbf{11.81}$	$\log p(s_2 \text{ELECTRA}) = -162.62$	0
s_1 : And diplomacy is more people to the place.	$\log p(s_1 \text{ELECTRA (PLL)}) = -62.81$	$\log p(s_1 \text{ELECTRA}) = -\mathbf{47.33}$	30
s_2 : Brezhnev ingenuity disembarking Acapulco methamphetamine arthropods unaccompanied Khrushchev.	$\log p(s_2 \text{ELECTRA (PLL)}) = -\mathbf{34.00}$	$\log p(s_2 \text{ELECTRA}) = -230.97$	0
s_1 : Sometimes what looks and feels real to you.	$\log p(s_1 \text{RoBERTa (PLL)}) = -36.58$	$\log p(s_1 \text{RoBERTa}) = -\mathbf{51.61}$	30
s_2 : Buying something breathes or crawls aesthetically to decorate.	$\log p(s_2 \text{RoBERTa (PLL)}) = -\mathbf{9.78}$	$\log p(s_2 \text{RoBERTa}) = -110.27$	0
s_1 : In most other high priority packages were affected.	$\log p(s_1 \text{RoBERTa (PLL)}) = -71.13$	$\log p(s_1 \text{RoBERTa}) = -\mathbf{61.60}$	30
s_2 : Stravinsky cupboard nanny contented burglar babysitting unsupervised bathtub.	$\log p(s_2 \text{RoBERTa (PLL)}) = -\mathbf{21.86}$	$\log p(s_2 \text{RoBERTa}) = -164.70$	0

对于每个双向模型，该表显示两个句子对，当其预测基于伪对数似然 (PLL) 估计时，模型严重失败。在这些句子对中的每一个中，PLL 估计有利于句子 s (较高的 PLL 粗体)，而近似对数似然估计和大多数呈现该句子对的人类受试者更喜欢句子 s_o 。（当多个句子对在模型中引起相同的最大误差时，表中包含的示例是随机选择的。）具有长、多标记单词（例如“甲基苯丙胺”）的句子具有较高的 PLL 估计，因为每个他们的代币可以被其他代币很好地预测。然而，根据人类的判断和基于适当的条件概率链的近似对数概率估计，整个句子是不可能的。

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
<input type="checkbox"/>	<input checked="" type="checkbox"/> The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
<input type="checkbox"/>	<input checked="" type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
<input type="checkbox"/>	<input checked="" type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided <i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>
<input type="checkbox"/>	<input checked="" type="checkbox"/> A description of all covariates tested
<input type="checkbox"/>	<input checked="" type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
<input type="checkbox"/>	<input checked="" type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
<input checked="" type="checkbox"/>	<input type="checkbox"/> For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted <i>Give P values as exact values whenever suitable.</i>
<input checked="" type="checkbox"/>	<input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
<input checked="" type="checkbox"/>	<input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
<input type="checkbox"/>	<input checked="" type="checkbox"/> Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The behavioral data was collected using a custom Gorilla script (https://gorilla.sc/). The Gorilla code will be provided upon request.
Data analysis	Data were analyzed with custom Python code employing the pandas and statsmodels libraries. Our complete analysis code is shared online at https://github.com/dpmlab/contstimlang. The repository also includes the code necessary for generating the controversial sentence pairs.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The experimental stimuli, detailed behavioral testing results, and code for reproducing all analyses and figures are available at https://github.com/dpmlab/contstimlang{github.com/dpmlab/contstimlang}.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Quantitative experimental study.
Research sample	Native English speaking US-based prolific.co users. 55 males and 45 females. The average participant age was 34.08 ± 12.32 . We chose to sample US-based native English speakers since this is the reference population for most of the language models we considered. Since the participants were free to decide whether to participate in our study, the sample is not necessarily representative. A follow up experiment (presented in Extended Data Figure 5 and described in detail in supplementary section 1.2) recruited additional 30 participants from the same sampling frame.
Sampling strategy	No formal subject sampling was employed (online requirement continued until the study was completed). The number of subjects was set before data collection according to our experience with similar designs (e.g., Golan, Raju & Kriegeskorte, 2020 PNAS).
Data collection	Data was collected online at the privacy of the participants' homes. The researchers were not present or involved in the experimental sessions (i.e., the behavioral experiment was fully automated).
Timing	June 14 2021 through August 8 2021. The follow up experiment was conducted from October 25 2022 through November 11 2022.
Data exclusions	We used a pre-established exclusion criteria to ensure that all analyzed participants showed sincere effort in their linguistic judgments. We included 12 trials with pairs of a naturally occurring sentences and their shuffled versions. We rejected the data of 21 participants who failed to choose the original, unshuffled sentence in at least 11 of the 12 control trials, and acquired data from 21 alternative participants instead, all of whom passed this data-quality threshold. Using the same criteria, we rejected the data of 3 participants in the follow up experiment.
Non-participation	5 participants failed to complete the main experiment.
Randomization	In the main experiment, we randomly allocated participants to replication groups (every ten subjects were presented with a different stimulus set).

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	See above
----------------------------	-----------

Recruitment

Participants were recruited through the Prolific.co website. We did not identify a particular self-selection bias that is likely to impact the results.

Ethics oversight

The Columbia University Institutional Review Board (protocol number IRB-AAAS0252).

Note that full information on the approval of the study protocol must also be provided in the manuscript.