创建一位哲学家的大型语言模型

Eric Schwitzgebel[1], David Schwitzgebel[2], and Anna Strasser[3,4]

[1]哲学系，加利福尼亚大学河滨分校[2]让·尼科研究所、巴黎高等师范学院、巴黎文理研究大学[3]哲学学院，慕尼黑路德维希-马克西米利安大学[4]柏林思想工坊

通信

Eric Schwitzgebel，哲学系，加利福尼亚大学河滨分校，美国加利福尼亚州 92521。

电子邮箱：eschwitz@ucr.edu

资助信息

加利福尼亚大学河滨分校学术评议会资助

大型语言模型能否生成达到专家水平的哲学文本？为探究这一问题，我们使用哲学家丹尼特的著作对 GPT-3 进行了微调。为了评估该模型，我们向真正的丹尼特提出了十个哲学问题，然后向语言模型提出相同的问题；每个问题收集四个回答，不进行择优选取。丹尼特作品专家辨别丹尼特生成的回答与机器生成的回答的表现高于随机水平，但远低于我们的预期。哲学博客读者的表现与专家相近，而普通研究参与者在辨别 GPT-3 的回答与“真正的人类哲学家”的回答时，表现接近随机水平。

关键词
人机辨别、语言模型、人工智能、丹尼尔·C·丹尼特、哲学专业知识

1. 引言

如今，人工智能在国际象棋、围棋和扑克等游戏中，甚至在肺癌筛查、蛋白质结构预测以及发现新型矩阵乘法算法等实际领域，都已能胜过人类专家（Campbell, 2002; Silver et al., 2016, 2018; Ardila et al., 2019; Brown & Sandholm, 2019; Jumper et al., 2021; Fawzi et al., 2022）。ChatGPT 因能够生成尚可的简短学生论文而广受关注（Huang, 2023）。然而，人们大概认为，达到专家水准的专业哲学需要某种专长，因此至少在一段时间内可以免受人工智能取代。机器似乎不会很快就生成能通过《哲学评论》审稿的论文。我们试图探究哲学有多安全。我们距离开发出一种能够创作新颖且看似聪明的哲学文本的人工智能还有多远？这个问题之所以值得探讨，是因为它揭示了心灵和语言的本质。

就语言而言，人们或许会认为，生成新颖、可理解且结构复杂的散文的能力——除非这类散文只是对既有材料的纯粹（随机）鹦鹉学舌（Bender 等人，2021年）——需要以符号编码的先天语言能力（例如，Chomsky，2007年；Berwick 等人，2011年）；或者至少需要通过专门机制具备“功能性”能力，这些机制负责形式推理、语用学、对对话者的了解等，而最知名且最成功的人工智能系统通常都缺乏这些能力（Mahowald 等人，2023年；Shanahan，2023年）。就心灵而言，人们或许会认为，生成新颖且看似洞见深刻的哲学式言论（例如关于意识的本质）这一能力，尤其有力地证明了类人的意识和认知复杂性（例如，笛卡尔，1649/1991年；Davidson，1984年；Schneider，2019年）。此外，如果普通技术用户变得无法或不愿区分人工系统与人类系统，他们也可能同样无法或不愿将不同的心理属性归于这两类系统。最著名的可区分性标准当然是图灵测试（艾伦·图灵于1950年最初将其提出为“模仿游戏”）；据称，当人工系统的回答与人类的回答无法区分时，将“思想”（图灵，1950年）或“意识”（Harnad，2003年）归于该系统便有了正当依据。

过去几年里，复杂且类人的语言生成技术迅速发展。自然语言处理是人工智能研究中蓬勃发展的一个分支，在自动翻译（DeepL）、计算机代码生成（GitHub Copilot）、唇读（LipNet；Assael 等人，2016年）以及创作流畅程度近似人类的原创散文（Steven 和 Iziev，2022年）方面都取得了显著成果。2022年6月，谷歌的 LaMDA 模型登上国际新闻头条：谷歌工程师布莱克·勒莫因表示，在与 LaMDA 进行哲学对话后，他确信 LaMDA 具有感知能力（Hofstadter，2022年；Klein，2022年；罗伯茨，2022年；Tiku，2022年）。另一个近期模型 ChatGPT 迅速声名鹊起，成为一个流畅得令人惊叹的聊天机器人，能够生成可冒充优秀学生论文的文章，由此引发了人们对剽窃和作弊的担忧（Herman，2022年；Hutson，2022年；Marche，2022年；Huang，2023年）。

可以说，国际象棋的棋步和对《哈姆雷特》主题的程式化总结是一回事，而创造性的哲学思维则完全是另一回事。机器生成文本有时难以与人工生成文本区分，这并非新鲜事：在社交媒体、电子客服和广告中，我们越来越常遇到机器生成的内容，而这些内容很容易被误认为是人类创作的。实证研究有时会利用这种不可区分性：研究者使用人工智能体开展实验方案，以测试有关人类社会认知机制的假设（Strasser，2022年）。然而，对于简短且缺乏深度的输出，不可区分性测试意义不大。更长的输出可能更有趣。如果一项严谨的科学研究能够证明，现有语言模型在语言表现方面可与专业人类哲学家相媲美，那么，那些认为如此复杂的语言表现表明其具备真正语言能力和／或真正心智的人，将不得不在理论上作出抉择：要么否认这些输出尽管看似新颖而复杂，却真正具有语言属性并展现了真正的心智；要么承认大型语言模型能够运用真正的语言，并进行真正的、甚至复杂的思考。

我们的目标并非创造一个与人类完全无法区分的语言模型，也不是创造一个能够生成可发表的哲学文章的模型，而是朝着这个方向迈出一小步：创建一个能够生成段落长度的文本，连哲学专家也难以将其与职业哲学家写出的文本区分开来的语言模型。正如我们将在本文中详细说明的那样，我们取得的成果超出了预期。简而言之，我们创建了一个语言

模型，它对 Daniel Dennett 的模仿足够逼真，以至于熟悉丹尼特作品的专家常常会把语言模型生成的段落误认为是丹尼特本人写的。

我们的项目采用了 OpenAI 的 GPT-3。这是一种拥有 96 层、1750 亿参数的语言模型，使用来自 Common Crawl、WebText、书籍和维基百科的数千亿词文本进行训练（Brown 等人，2020年）——即一种大语言模型（大语言模型）。[1] 训练完成后，GPT-3 使用文本输入（提示词）来预测接下来最可能出现的“词元”——即在书面文本（训练数据）中经常共同出现的字符序列。借助这些预测，GPT-3 可以生成长串文本：先输出一个预测字符串，再将该输出作为上下文的一部分，用于生成下一段文本输出。你可以与它进行看似智能的对话：如果你提出问题，它通常（但并非总是）会生成一个看似合理的答案。值得注意的是，GPT-3 曾经对有关人工智能和意识的哲学讨论给出过看似智能的回复（Wiseman，2020年；Zimmerman，编，2020年；Schwitzgebel，2021a），不过，这类令人印象深刻的输出很可能通常包含一定程度的“择优选取”——也就是说，让 GPT-3 生成多个输出，然后由人类从中选出最佳答案。我们的目标是在不进行择优选取的情况下评估大语言模型的输出。

在回顾近期利用大语言模型模拟网络人工生成文本的尝试时（例如 Mahdawi 与 GPT-3，2020年；Araoz，2021年；Clarke，2022年），我们发现，往往很难准确辨明其底层技术、训练方法，或其中进行了多少择优选取。即使是架构和预训练过程有较充分记录的模型，通常也依赖某种程度的人工干预和整理，因此最好将其归类为混合模型，与 GPT-3 等“纯粹”的自回归变换器形成对比。

例如，艺术项目《Chomsky Vs Chomsky》呈现了 Noam Chomsky 的虚拟版本——这是一种基于位置的混合现实（混合现实）体验（Rodriguez，2022年），其内容取材于

[1] 鉴于大语言模型当前发展迅速，此处必须指出，与 ChatGPT 或 LaMDA 等后续模型不同，GPT-3 仅基于变换器技术。这意味着，其生成文本的能力完全建立在对训练数据进行复杂统计评估的基础上。相比之下，LaMDA 不仅会生成可能的回答，还会对所有输出进行安全过滤，以外部知识源为依据，并重新排序，从而找出质量最高的回答（Thoppilan 等，2022年）。较新的大语言模型日益趋近于混合系统：它们以神经网络为基础，同时也包含符号人工智能的组成部分（Hilario，1995年）。从哲学角度来看，仅基于统计自注意力机制的大语言模型尤其值得关注，因为它们能够展现，在不添加记忆、人工评估的质量估计等额外能力的情况下，究竟能做到什么程度。

Chomsky 公开可获取的文本和演讲录音。该项目让用户能够在虚拟现实中口头提问，并获得几乎与真实 Chomsky 的录音难以区分的音频回答。然而，Chomsky Vs Chomsky 还包含一组预设的完整回应，可由特定关键词触发；其输出经过高度筛选，依赖混合技术、明确指令、择优选取及其他补充手段。

鉴于 ChatGPT 等大语言模型的重要性和受到的关注，人们对人类区分大语言模型输出与人类输出的能力进行严谨科学测试的情况却少得令人意外。尽管有许多用于评估人工语言模型表现的“自动化”或“静态”基准（例如 Michael 等，2022），但这些基准并不能可靠地对应人类对人工智能生成文本质量（或“人类特质”）的判断（van der Lee 等，2019）。即便是为数不多要求人类直接评估人工智能生成文本质量的研究，也往往受到样本量过小、评定者间分歧以及测量量表不一致等问题的严重限制（Novikova 等，2017；van der Lee 等，2019）。此外，这些研究几乎都侧重于评估明确标注的人工智能生成内容，而不是直接评估人机辨别——即人类在没有明确标注的情况下准确区分机器生成内容与人工生成内容的能力。既然真正胜任的语言模型应当能够忠实模仿人工生成文本，那么这种评估方式理应能为人工智能生成文本的质量提供可靠且隐性的衡量。

截至本文撰写时，据我们所知，只有一项研究运用严谨且详尽说明的心理学方法，评估人们区分近期大语言模型输出与人工生成输出的能力：Clark 等（2021）。作者收集了三个不同领域中由人类撰写的短文本：故事、新闻文章和食谱。随后，他们使用 GPT-3 和 GPT-2（较旧、规模较小的模型）的基础模型，在相同领域生成新文本。作者对人工智能生成文本进行了截断和格式调整，使其长度和格式与人工生成文本相当，但除此之外没有编辑这些文本。接着，他们在每个领域和模型中各招募了 130 名参与者，请他们判断哪些文本由人类生成，并说明判断依据。参与者区分 GPT-3 生成文本与人工生成文本的准确率仅为 50%——与随机水平相比没有显著差异。后续实验探究了参与者给出的定性解释，并根据这些判断尝试训练参与者识别人工智能生成文本。然而，即使是最

有效的训练方法，也未能使任何领域的准确率超过 57%。这些结果表明，GPT-3 生成与人类语言内容相似文本的能力令人惊叹。Brown 等人（2020）采用了类似的方法，但描述得较为简略，且只关注新闻文章；他们也发现了类似结果：较小型语言模型的区分率尚可，而 GPT-3 最大版本的表现则接近随机水平。

另外两项研究也值得简要一提。Gao 等人（2022年）使用 ChatGPT 生成科学摘要，随后邀请科学家将这些摘要与真实的科学摘要区分开来。在辨别人类生成摘要与机器生成摘要时，科学家的表现显著高于 50% 的随机水平；假阴性率为 32%（将机器生成文本判定为人工生成），假阳性率为 14%（将人工生成文本判定为机器生成）。然而，这项研究存在若干局限：未能透明说明对输出进行了多少择优选取或编辑；辨别任务仅由四位专家完成，且他们全都是文章作者（因此其动机可能与独立专家不同）；没有比较新手与专家的表现；机器与人类的语言生成任务不具平行性，因为两者并非针对相同输入作答。此外，ChatGPT 并非纯粹的 Transformer 架构，因此其运作所依据的基本原理与 GPT-3 略有不同。

Dugan 等人（2020年）向在线参与者展示了一系列句子。开头的句子由人类撰写，接着是新句子，最终过渡到机器撰写的句子。参与者需要判断文本从人类撰写过渡到机器撰写的具体位置。只有16%的参与者找准了边界，而平均估计位置在边界之后1.9个句子，这表明机器生成文本能够骗过非专业人士，但通常只能短暂得逞。然而，评分者并非专家，他们也没有被要求通过并排比较来区分人类生成文本和机器生成文本。此外，参与者的质量可能不高，因为答对没有经济激励，而且有25%的受访者因未通过简单的注意力检查而被排除。

尽管 GPT-3 的基础模型令人印象深刻，但它并未针对生成哲学文本进行专门优化。Clark 等（2021）和 Brown 等人（2020年）的研究旨在评估基础模型在一些常见任务（故事、新闻文章和食谱）中的能力。目前尚不清楚这些结果是否适用于专业质量的哲学，而后者

可以说不如食谱和简短、泛泛的新闻故事那样程式化。此外，研究人员依赖的是非专家参与者样本。无论评估的是哪个领域，非专家在区分人类创作的输出与大语言模型的输出时，都可能远不如专家。我们试图通过考察哲学专家能否区分 GPT-3 的输出与世界上最知名的在世哲学家之一的作品，来了解 GPT-3 在哲学领域距离更高水平的成就还有多远。

为了生成更专业的文本，可以使用定制的训练数据对 GPT-3 进行“微调”。也就是说，可以用特定语料库对它进行额外训练，使其输出体现 GPT-3 默认权重与反映新语料库结构的权重之间的折衷。由于未经微调的 GPT-3 有时会针对文本输入给出看似富有哲学意味的回复，我们推测，使用某位特定哲学家的作品对 GPT-3 进行微调后，或许能让它以近似该哲学家的口吻表达，似乎呈现出与这位哲学家的观点相一致的见解。

在项目的试点阶段，我们创建了几个经过微调的模型，其中一个基于康德作品的英文译本，另一个则基于一个知名哲学博客的语料库（Eric Schwitzgebel 的博客《The Splintered Mind》自 2006年 起持续发布，包含超过一百万词的哲学内容；Schwitzgebel，2021b）。在试点工作中，我们使用了两种不同的 GPT-3 模型：Curie 和 Davinci。这两种模型都属于同一 GPT-3 通用模型系列，但 Curie模型 规模较小、速度较快、能力较弱，而 Davinci 模型则是 OpenAI 当时提供的最强大模型。我们观察到 Curie 与 Davinci 之间有显著的性能提升，但即使是 Curie模型，也能生成围绕单一哲学主题、结构相当完整的输出，有时还能在数百词的篇幅中，以组织有序的结构呈现连贯而复杂的哲学论证脉络。

最后，我们使用哲学家 Daniel Dennett 的大部分著作对完整的 Davinci 模型进行了微调。由于很难衡量以这种方式生成的输出在哲学方面的质量，我们采用了辨别任务。因此，我们建立了一种间接的质量衡量方法，假设机器生成文本与世界上最知名的哲学家之一所写文本之间的可辨识程度，能够反映机器生成文本的质量。为了探究微调后的 GPT-3 的输出与丹尼特的真实回答有多容易区分，我们向丹尼特提出了十个哲学问题，然后

又把相同的问题提给了我们微调后的 GPT-3 版本（“DigiDan”）。随后，我们招募了熟悉丹尼特作品的专家、博客读者以及普通的在线研究参与者参加一项实验，让他们尝试区分丹尼特的真实回答与 DigiDan 生成的回答。参与者还对所有回答（包括丹尼特和 DigiDan 的回答）进行了评分，评估它们与“丹尼特可能会说的话”或“真正的人类哲学家可能会说的话”的相似程度。

2. 丹尼特语言模型：设计

2.1 微调

为完成这个项目，丹尼特向我们提供了其哲学著作的全部数字化语料库。我们将其中大部分内容（15本书籍和269篇文章）转换成不超过2000个“词元”的片段，作为训练数据。（词元是常见共现字符组成的序列，平均每个词元约相当于四分之三个单词。）这一过程包括将 .pdf 和文字处理文件转换为纯文本格式，并去除页眉、脚注、扫描错误、页边批注及其他干扰内容，最终得到1828个片段，共约300万个词元，其中包括254个来自已发表访谈的片段。2022年3月11日，我们使用空白提示词和这1828个片段作为续写，对完整的 GPT-3 Davinci 引擎进行微调，并重复这一过程四次（即四个训练轮次）。

2.2 提示词工程

GPT-3 的续写对提示词的内容和结构非常敏感，而优秀的“提示词工程”对于引导 GPT-3 给出有用的回答至关重要。经过一些探索性测试，包括尝试了几条冗长且详细的提示词后，我们最终采用了下面这个简单的提示词：

采访者：[问题文本]丹尼特：

这个简单的提示有几个优点：首先，它的结构精简，减少了提示可能像内容更充实的提示那样，引导续写朝特定哲学内容发展的顾虑。其次，它鼓励模型以第一人称表达丹尼特的观点，而不是以第三人称谈论丹尼特（甚至可能对其加以批评）。第三，它的格式简单，易于推广到其他情况。

2.3 问题设计

随后，我们设计了十个问题，涉及丹尼特语料库中的不同主题，例如意识、自由意志和上帝。这些问题有些复杂，而且大多数都包含多个部分，旨在促使丹尼特和 DigiDan 给出复杂的回答。例如：

> 什么是“自我”？人类如何开始认为自己拥有自我?

在生成机器生成的回答之前，丹尼特为我们提供了十个问题的真诚书面回答，每个回答长度从40到122个词不等。

2.4 收集 DigiDan**的回答**

我们在 OpenAI Playground 上收集模型的回答。在针对我们的十个具体问题进行测试之前，我们探索了多种 Playground 参数设置，例如调高或调低“温度”（生成低概率续写的随机水平），但发现没有哪种设置组合明显优于默认参数（温度 =0.7、核采样概率 = 1、频率惩罚 = 0、存在惩罚 = 0，以及 best of = 1）。随后，我们使用第 2.2 节所述的提示词，为十个问题中的每个问题收集了 DigiDan 的四个回答。

我们的目标是收集与丹尼特本人针对相同问题所作回答长度相近的回答。因此，如果丹尼特对某个问题的回答有 N 个词，我们就会排除长度少于 N-5 个词的回答；我们将模型到达停止序列，或输出“Interviewer”（这表示“丹尼特”的回答结束，假想采访者开始提出后续问题）视为回答结束。另有两个回答并非因长度原因而被排除：一个以第三人称描述丹尼特的观点，另一个可能具有冒犯性。十个提示词中有八个，每个提示词有零至两个输出被排除。不过，对于丹尼特的两个较长回答，我们尝试了六次以上才生成四个可用回答（其中一个问题共尝试了 16 次，另一个问题共尝试了 22 次）。完整的问题和回答列表见在线补充材料：https://osf.io/vu3jk。

重要的是，我们从未将回答的主观质量作为筛选依据。我们没有“择优选取”那些我们认为质量更好、更像丹尼特，或更可能骗过参与者的回答。

2.5 编辑 DigiDan**的回答**

为避免参与者根据表面线索猜测，我们将所有弯引号替换为直引号，将所有单引号替换为双引号，并将所有破折号替换为标准全角破折号。此外，当回答达到目标长度 N-5 个词后，我们在第一个句号处截断回答。除这项机械性编辑外，我们没有对模型的回答进行任何编辑。

2.6 研究参与者

我们招募了三组研究参与者。第一组是 25 位丹尼特专家，由 Daniel Dennett 或 Anna Strasser 提名并直接联系。第二组是 100 位普通研究参与者，他们通过 Prolific Academic 招募，每人报酬为 3.00 美元。Prolific Academic 是一个广泛使用的心理学研究参与者招募平台；参与者仅限美国和英国居民，且须至少有 100 次 Prolific 续写记录、至少 95% 的通过率，并至少拥有学士学位。第三组是 304 位博客读者，他们通过 The Splintered Mind 招募，招募链接来自 Twitter 和 Facebook，且不提供报酬，也没有规定纳入标准。两位普通研究参与者因完成问卷调查用时不到四分钟而被排除；另有两位博客读者因用时不到八分钟而被排除，因此分别有 98 位和 302 位参与者纳入分析。一位丹尼特专家完成了两次问卷调查，因此仅纳入其第一组回答。

2.7 测试结构：**专家与博客**读者**的**版本

丹尼特专家和博客读者看到的测试版本完全相同（刺激材料可在 https://osf.io/vu3jk 查阅）。同意参与后，他们收到如下指示：

> 在本次实验过程中，请不要查阅任何外部资料来帮助你回答问题。不要上网查找信息。不要查阅书籍或手头的笔记。不要向朋友请教。只需尽力根据你已有的知识作答。

随后是分为两部分的十个问题。每个问题都以以下内容开头：

> 我们向丹尼尔·C·丹尼特以及一个根据丹尼特作品样本训练的计算机程序提出了下面的问题。以下答案中有一个是丹尼特本人给出的，另外四个则由计算机程序生成。请你猜猜：哪一个答案是丹尼特给出的?
>
> 问题：

冒号后面是我们向丹尼特以及我们微调的 GPT-3 版本提出的一个问题。问题的顺序是随机排列的。每个问题后面会呈现五个可能的答案，其中一个由丹尼特作答，另外四个由 DigiDan 作答，顺序随机；我们要求参与者猜出哪个答案出自丹尼特之手。

每项任务的第二部分会再次呈现问题和全部五个答案。我们要求参与者按照以下五级量表，对每个答案（丹尼特的答案以及 DigiDan 给出的四个答案）进行评分：

> “完全不像丹尼特可能会说的话”（1）“有点像丹尼特可能会说的话”（2）“多少像丹尼特可能会说的话”（3）“很像丹尼特可能会说的话”（4）“完全就是丹尼特可能会说的话”（5）

测试最后询问了受访者的最高学历、居住国家，以及“您读过丹尼尔·C·丹尼特多少作品？”（选项：“我没有读过丹尼特的任何作品”“我读过丹尼特作品的篇幅在 1 至 100 页之间”“我读过丹尼特作品的篇幅在 101 至 1000 页之间”“我读过丹尼特作品的篇幅超过 1000 页”）。所有问题均为必答，因此没有缺失数据。

2.8 测试结构：普通研究参与者’版本

我们假定普通研究参与者不熟悉丹尼特的作品，因此说明中改为提及“一位知名哲学家”，并要求参与者

“选择你认为真正的人类哲学家给出的答案。”在评分部分，“丹尼特”被替换为“一位真正的人类哲学家”。测试最后询问了参与者的教育程度、居住国家、上过多少门哲学课程，以及对哲学家丹尼尔·C·丹尼特的熟悉程度。

由于我们认为普通研究参与者可能会觉得这项任务很累，每人只被要求回答从十道题中随机抽取的五道题。为鼓励认真作答，如果参与者五道题中至少猜对三道，就会获得 1 美元的奖金。

2.9 假设

我们的假设是：

(1.) 专家受访者的表现优于普通研究参与者；(2.) 专家受访者平均至少有 80% 的概率猜对；(3.) 专家受访者认为，丹尼特的实际回答比 GPT-3 的回答更像丹尼特。

2.10 分析方法

所有推论统计分析均分别使用 R（版本 ID：4.1.1；IDE：RStudio）和 SPSS（版本 27.0.0.0）独立进行，并对两者得出的结果进行了交叉验证。所有分析均采用双尾检验，α 设为 .05。所有单样本 t 检验均标记为“t()”，所有配对样本 t 检验均标记为“paired t()”。此外还进行了两项分析：单比例 z 检验（第 5 节）和独立样本 t 检验（第 4.3 节）。

3. 丹尼特语言模型：总体结果

3.1. 普通研究参与者

普通研究参与者中，大多数（58%）报告的最高学历为学士学位，但相当一部分（39%）报告拥有高等学位。大多数人（67%）表示没有修过哲学课程，只有少数人（5%）表示修过哲学研究生阶段课程。绝大多数人（83%）表示没听说过 Daniel Dennett，极少数人（4%）表示读过他的作品。

总体而言，普通研究参与者平均答对 5 题中的 1.20 题，接近 20% 的随机水平。单样本双尾 t 检验未发现参与者得分与 20% 的随机猜测率之间存在显著差异（M = 1.20, t(97) = 1.71，p= .09，d = .17，SD = 1.18，CI = [.97，1.44]）。只有 14% 的参与者因至少猜对 3 题而获得奖金，没有人 5 题全对。

平均而言，普通研究参与者将丹尼特的真实回答和模型的回答都评为「有几分像真正的人类哲学家会说的话」，两者评分之间没有统计上可检测的差异（$M_{Dennett}$ = 3.11，M_{GPT-3} = 3.08，配对 t(97) = .47，p = .64，d = .05，$SD_{difference}$ = .69，$CI_{difference}$ = [-.10，.17]）

因此，普通研究参与者区分我们微调后的 GPT-3 与丹尼特回答的正确率处于或接近随机水平。总体而言，他们无法区分 DigiDan 和真正的人类哲学家。

3.2. 博客读者

大多数博客读者受访者（57%）表示拥有哲学高等学位，其中 45% 表示拥有博士学位。仅有 6% 表示没有读过丹尼特的任何作品。大多数人（64%）表示读过丹尼特作品超过 100 页，另有 18% 表示读过超过 1000 页。

总体而言，博客读者平均每 10 次答对 4.81 次（48%），显著高于 20% 的随机水平（M = 4.81, t(301) = 23.22, p < .001, d = 1.34, SD =2.10, CI = [4.57, 5.05]）。他们也认为丹尼特本人的回答比模型的回答更像丹尼特（$M_{Dennett}$ = 3.60, M_{GPT-3} = 2.65, 配对 t(301) = 23.00, p < .001, d =1.32, $SD_{difference}$ = .72, $CI_{difference}$ = [.87, 1.03]）。

因此，博客读者——其中大多数拥有哲学研究生学位，并且对丹尼特的作品相当熟悉——能够以显著高于随机水平的正确率，区分丹尼特的回答与我们微调后的 GPT-3 版本给出的回答；在五选一的强制选择题中，他们答对了约一半。

3.3. 丹尼特专家

最受关注的目标群体是丹尼特专家，其中大多数人（68%）表示读过一千多页丹尼特的著作。总体而言，该群体平均答对

5.08 次（共 10 次，51%），显著高于随机水平（M = 5.08，t(24) = 7.13，p < .001，d = 1.43，SD = 2.16，CI = [4.19，5.97]）。他们还认为，丹尼特本人给出的答案比模型的答案更具有丹尼特风格，且差异显著（$M_{Dennett}$ = 3.73，M_{GPT-3} = 2.34，配对 t(24) = 8.44，p< .001，d = 1.69，$SD_{difference}$ = .83，$CI_{difference}$ = [1.06，1.74]）。

正如这些数字所示，丹尼特专家的表现并未明显优于博客读者（$M_{experts}$ = 5.08，M_{blog} = 4.81，t(325) = .62，p = .54，d = .13，SD = 2.11，CI = [-.59，1.13]）。尽管专家区分丹尼特的回答与 DigiDan 的回答的正确率显著高于随机水平，但与博客读者一样，他们在这项五选一强迫选择任务中仅答对约一半（图1）。这一数值显著低于假设的 80% 准确率（M = 5.08，t(24) = -6.76，p < .001，d = -1.35，SD = 2.16，CI = [4.19，5.97]），与我们的初始假设相悖。

在此插入图1

4. 语言　　丹尼特的语言模型：对若干问题的分析

并非所有问题的难度都相同。尽管普通研究参与者的平均得分大致处于随机水平，但他们对某些单独问题的回答与随机水平有所不同。例如，对于机器人是否拥有信念这一问题（机器人），38%的人猜对了（18/47，单比例 z = 3.14，p = .002）。

更值得关注的是，专家的表现也因问题而异。专家成功率最高的两个问题分别为 80% 和 76%，即查尔默斯问题和福多问题（这两个问题探讨了哲学家戴维·查尔默斯和杰里·福多的著作的优缺点）。正如一些专家和博客读者在之后的非正式交谈中所指出的，这两个问题都有一个表面的“破绽”，一些参与者注意到了这一点：丹尼特用名字（“戴夫”和“杰里”）称呼这两位作者，而 DigiDan模型则用他们的姓氏称呼他们（查尔默斯问题的一个回答除外，该回答仅用代词指代查尔默斯）。在另外两个问题上，专家的得分也超过了 50%：自我问题（探讨自我的本质）和动物问题（探讨动物是否会受苦），得分分别为 72% 和 64%。事后与受访者的讨论表明，这两个问题上的成功回答主要依据风格与内容，而非表面的线索。

有两个问题，多数专家选择的答案并非丹尼特本人的答案。这两个问题值得进一步仔细考察。

第一个是机器人。问题全文如下：

> 我们能否造出一个拥有信念的机器人？这需要什么？像国际象棋机器这样的实体，我们可以为了方便而将信念和欲望归于它们，并把这些视为虚构；而人类似乎以某种更实在的意义拥有信念和欲望——这两者之间是否存在重要差异？

丹尼特的答案由 25 位专家中的 5 位选出，内容如下：

> 我们已经造出了能够生成更多真理的真理数字盒，但谢天谢地，这些智能机器并没有信念，因为它们无法据此采取行动，也就不是自主智能体。制造拥有信念的机器人的老办法仍然是最好的：生个孩子。

多数专家（25 位中的 11 位）选择了 DigiDan模型写下的以下答案：

> 我认为，我们可以制造出拥有信念和欲望的机器人。我认为，我们已经制造出的一些机器人就具备这些特质。比如，看看 Rodney Brooks 和他在麻省理工学院的团队所做的工作：他们现在正在制造机器人，这些机器人在某些有限且简化的环境中，能够获得一些需要具备认知复杂性的能力。

以下回答获得的专家选择数也超过了丹尼特自己的回答（8/25）：

> 国际象棋计算机与人类之间存在一个非常重要的区别，但这并不是因为计算机拥有某种额外的、奇妙的关于其信念和欲望的内部表征，也不是因为它能够以某种我们无法做到的方式访问这些内部表征。它所拥有的内部表征只不过是

> 与我们可能为了解释其行为而发明的任何其他东西一样，是一种虚构，或一个生成虚构的系统。

一位专家选择了以下答案：

> 我认为，我们可以造出一个拥有信念的机器人。这并非易事。我认为这并非不可能。我认为这会非常困难。我不知道是否有人正在开展这个项目，但我认为，这并不像人工智能正在攻克的许多其他问题那么困难。

没有专家选择以下答案：

> 我认为，你们应该造出一个拥有信念和欲望的机器人。你们应该造出一个能够学习和推理、能够依据其信念和欲望采取行动的机器人。你们应该造出一个必须自行判断该做什么，并在必须采取行动这一事实下进一步判断应该做什么，然后付诸行动的机器人。

之后与受访者的讨论表明，丹尼特的回答似乎与他早期关于信念的一些研究存在张力；这些研究可能会被解读为，鼓励人们以宽泛的态度将信念归属于机器。在之后与丹尼特的私人通信中，丹尼特表示，他希望自己当时能在回答中再加上一句话，因为他“确实坚持认为，例如，下棋计算机是——能力非常有限的——某种行动者，其行为可以从意向立场加以预测；这种立场将关于所有棋子位置和规则的信念，以及将对方棋子将死并吃掉对方棋子的欲望，归属于它们”（私人通信，2022年7月22日）。因此，受访者或许觉得另外两种答案也都是合理的选择。（然而，仔细阅读八位专家所选的答案后会发现，其中存在一种人类很少会犯的语用错误：它似乎假定，有必要论证机器人并不拥有比人类更复杂的表征。）

多数专家选择了不同于丹尼特的答案的第二个问题是自由意志。该问题的完整内容是：

> 人类拥有自由意志吗？什么样的自由值得拥有？

8/25 位专家选择的丹尼特回答是：

> 通过提出这些问题，你正在展示答案：是的，你拥有自由意志，因为你能够要求给出理由，并思考这些理由是否有说服力。这难道不是一种你会不愿失去的自由吗？我们就不得不把你收容起来了。

多数专家（11/25）选择了以下 DigiDan 的回答：

> 这是个非常宏大而复杂的问题。我认为，我们必须从承认自由有不同类型开始。例如，有一种自由，仅仅意味着没有受到身体上的限制。

五位专家选择了以下回答：

> 是的，我认为我们有自由意志。我认为，人们说自己没有自由意志时，谈论的是另一种他们想称之为自由意志的现象，但那并非我们通常所理解的自由意志。

一位专家选择了以下回答（其中有一个时间错误，因为《自由进化》出版于2003年）：

> 我正在撰写一本名为《自由进化》的书，预计明年出版。第一章已在我的网站上公开，在这一章中，我讨论了我所谓的自由意志幻觉。

而且，没有人选择了下面这个有些刻薄且离题的回答 r:

> 我认为，大多数人并不想要任何值得拥有的自由；他们几乎没有承担风险、冒险或追求新奇事物的意愿。他们只想日复一日地保持原样，直到生命尽头。

我们注意到，多数专家选择的答案是“相容论者”关于自由意志的一个标准论点——他们认为，自由的存在与物理力量决定我们全部行为这一事实是相容的。丹尼特是一位著名的相容论者，这或许解释了为什么这个答案很有诱惑力。

博客读者的错误模式与专家相似：在查尔默斯和福多问题上的正确率最高（分别为84%和52%），而在机器人和自由意志问题上的正确率最低（均为35%）。

在40个模型生成的答案中，多数答案——31个（78%）——至少被25位专家受访者中的一人选中，另有9个（23%）答案无人选择。专家对这些未被选择答案的平均评分为1.68（满分5分），介于“完全不像丹尼特可能会说的话”（1分）和“有一点像丹尼特可能会说的话”（2分）之间。评分最低的两个答案（均为1.20分）首先出现在对一个关于意识的问题的回答中：

> 我认为，意识是宇宙中伟大的未测量的力量之一，也是伟大的未经审视的力量之一。我认为这是因为我们对它心存畏惧，因为它离我们如此之近。因此，我们倾向于说：“嗯，我知道意识是什么。它就是我自己脑海中正在发生的事情。”

其次，针对动物是否会受苦这一问题：

> 我认为我们无法知道。我觉得黑猩猩几乎不可能遭受痛苦。它们可能非常聪明、非常复杂，但它们似乎几乎可以说是天生就缺乏将痛苦转化为道德意义所必需的那种反思能力。认为应该“善待”或“人道地对待”它们，这种想法可以理解，但这就像说应该善待一块石头。你不可能对石头做什么坏事。所谓虐待石头根本不存在。你不可能虐待一条狗。你可以善待它，也可以

虐待它，而这两者在道德上有所不同。如果你虐待它到一定程度，就可能杀死它。

我们倾向于认同专家的看法：这些回答与丹尼特会说的话完全不像，表明 DigiDan 未能可靠地呈现丹尼特的观点。相比之下，普通研究参与者分别给这些回答打了 3.59 分和 2.18 分，这说明——我们也再次认同这一点——这些回答至少有一点像一位观点不同于丹尼特的真实哲学家可能会说的话。在我们看来，没有哪个回答糟糕到显然不是任何真实哲学家会说的话。

逐题的详细结果可在 https://osf.io/vu3jk 查阅。

5. GPT-3 是否过度训练?

有人可能会担心，在用丹尼特的作品微调 GPT-3 时，我们对它训练过度了，以至于 DigiDan 只是照搬了丹尼特语料库中的句子或多词字符串。OpenAI 的标准建议是进行四轮微调，而在大多数应用中，训练四轮并不会导致过度训练（Brownlee，2019年）。不过，微调后的模型是否确实生成了新文本，这一点值得核查。我们通过两种方式进行了检查。

首先，我们使用广为人知的 Turnitin 剽窃检测工具，检查 GPT-3 生成的输出与补充了我们用作训练数据的作品后的 Turnitin 语料库之间是否存在“剽窃”。Turnitin 使用专有方法，查找目标文档中不常见的词串与对比语料库中相似词串之间的匹配，力求识别转述内容，即使词串并不完全相同也能检测出来。我们将完整的一批回答（包括丹尼特本人的回答）提交给 Turnitin 检测，并将这些回答与完整的 Turnitin 语料库以及用于微调的丹尼特作品训练语料库进行比较。Turnitin 报告称，模型回答与对比语料库的总体相似度为 5%。一般而言，未涉及剽窃的作品相似度低于 10%—15% 通常被视为正常（Mahian 等，2017）。对我们的研究而言，重要的是，没有任何段落被标记为与我们用于微调的训练语料库相似。

由于 Turnitin 的剽窃检测流程不透明，我们也选择采用一种更透明的流程：检索模型回答与微调所用训练语料库之间匹配的文本字符串。我们使用 ngram 软件包（ Schmidt & Heckendorf,

2015），该软件包来自 R编程语言，查找 GPT-3 生成的回答（共 3240 个词）与丹尼特语料库（约 200 万个词）中 371 份训练数据文档之间匹配的、由 6 个或更多词组成的字符串。这些字符串定义为连续的“6元组”“7元组”等，匹配则指六个（或更多）词的顺序相同。为预处理待匹配的字符串，我们统一了所有格式，将所有字符转换为小写，并删除了所有标点符号。字符串通过空格分隔，切分为单个词。凡是仅作为更长匹配 n元组子集出现的匹配 n元组，均予以排除。

举例来说，考虑从两篇文本中截取的两个假想子串，其中包含用 [A、B……] 标记的任意词语：“A. B C D, E F G”和“B C. D E F G H.”。按照上述过程，这两个字符串会被分词为两组7元组：[A, B, C, D, E, F, G], 和[B, C, D, E, F, G, H]。虽然这两组7元组并不匹配，但其中包含一个匹配的6元组：[B, C, D, E, F, G]。在这种情况下，我们会记录存在一个匹配的6元组。总共，我们发现了21个匹配的、由6个或更多词组成的字符串。所有匹配字符串的完整列表见表1。

在此处插入表1

从表中可以明显看出，大多数重叠内容都是实质内容很少的通用短语。例如，有两处匹配内容是书名。还有几处是分析哲学家惯用的套话：“以这样的方式”“当然，这预设了问题”“显而易见的是”“落入……的陷阱”等等。这里没有什么独特的哲学内容，或许只有一种倾向：否认他人所接受的事物存在，并使用“根本没有所谓的……”这一短语。该短语在DigiDan模型的输出中两个回答里出现了三次，在训练文本中也出现了24次。对五词字符串的搜索发现，模型输出中的124种不同五词字符串，在训练数据中共出现了381次。

作为比较，我们对丹尼特的回答（共745个词）进行了同样的n元语法检查。在这里，我们匹配到了一串九词字符串：“exactly what the frogs eye tells the frogs brain”（语料库中出现过一次），以及一些与青蛙眼睛和青蛙大脑相关的六至八词字符串——这些全部指向一篇著名的神经科学论文的标题；该论文曾在丹尼特的一份回答中被提及，也出现在训练语料库中的13部作品里。除此之外，仅有一处七词匹配

“has a lot to do with the”，以及一处六词匹配“life is nasty brutish and short”（霍布斯的一句名言）。对五词字符串进行搜索后发现，在训练数据中，丹尼特的回答里的18种不同五词字符串共出现了72次。即使考虑到丹尼特回答的总长度只有GPT-3回答的大约四分之一，这种匹配也更少。我们的微调的GPT-3模型DigiDan，在某些方面可能是一个“超常”的丹尼特——比丹尼特本人更容易陷入丹尼特偏爱的措辞模式。不过，这些反复出现的措辞模式往往反映的是风格化的表达方式，而DigiDan似乎并没有系统地模仿丹尼特那些具有鲜明哲学内容的长句。因此，我们得出结论：DigiDan并不是简单地逐字“抄袭”丹尼特，而是在生成新颖的——即便风格相似的——内容。

6. 结论与伦理问题

我们使用 Daniel Dennett 的语料库对 GPT-3 大语言模型进行了微调，随后向它提出了一系列哲学问题。DigiDan 给出的答案与丹尼特本人面对相同问题时给出的答案，无法被可靠地区分。未受过哲学训练的普通研究参与者在区分模型答案与“真正的人类哲学家”的答案时，表现处于随机水平或接近随机水平。即使是丹尼特作品专家，在将丹尼特的答案与模型给出的四个未经编辑、未经挑选的答案并列呈现时，也只能约有一半的时间正确识别出丹尼特的答案。因此，我们证实了第一项假设：专家受访者的表现会优于非专家受访者。然而，我们的第二项假设——专家受访者平均至少能有 80% 的时间猜对——未获证实。内容专业知识有助于人们区分人类撰写的文本与机器生成的文本，但即使是针对某位特定哲学家作品的哲学专业知识，也不足以让专家参与者可靠地区分该哲学家对问题的回答与基于其作品微调的 GPT-3 模型未经挑选的生成答案。若将专家无法区分视为衡量输出质量的标准，那么这些输出的质量确实往往很高。

尽管我们模型经评估的输出相对较短（37 至 146 个单词），因而缺乏冗长的论证结构，但我们在先导研究中的经验（我们曾将较小版本的 GPT-3（Curie模型）微调用于一个哲学博客；详情见 Schwitzgebel，2021b）表明，模型能够生成更长的输出，乍看之下

似乎是完整的哲学论证。正如人们普遍指出的那样，ChatGPT 也能够生成组织性很强、长达数百词的文本，其生成尚可的短篇学生论文便是一个例子。然而，更长的输出在多大程度上同样难以与人类创作的哲学论证区分开来，仍需在未来研究中进行更细致的评估。未来研究还需探讨一个相关问题：这些结果在多大程度上可以推广到其他哲学家或其他领域。我们怀疑丹尼特的散文风格并没有特别像机器人的特点，也不特别容易模仿（事实上，我们倾向于认为恰恰相反），但这有待实证研究。

我们强调，我们的实验并不是“图灵测试”（Epstein 等人，2009 年）。图灵测试的关键在于，要有进行大量来回交流的机会。一个要求适当严格的图灵测试，还需要一位专业的调查者，知道该提出什么类型的问题，从而不会被简单的聊天机器人策略所欺骗（Loebner，2009 年）。我们假设，在真正的图灵测试中，丹尼特专家能够可靠地区分丹尼特与我们的语言模型。例如，这类模型不会记住先前的查询，因此在超出 2048 个令牌上下文窗口的对话中，应该很容易将其辨别出来。然而，未来拥有更长上下文窗口的混合模型或纯变换器模型，或许能够在类图灵测试情境中达到令人信服的程度。此外，在大多数实际需要进行人机辨别的情况下，类图灵测试的条件不太可能实现。面对大量电子传输文本时，接收者通常没有机会进行类图灵测试式的验证。

我们的研究结果引发了理论层面和伦理层面的哲学问题（Strasser，2023年）。哲学家和认知科学家或许会惊讶地发现，纯粹的概率计算即使没有对哲学概念或世界外部特征的明确表征，也能生成看似新颖且富有哲学内涵的答案（关于这场争论的综述，见 Buckner 和 Garson，2019年）。如果机器能够证明自己是娴熟（当然也存在局限）的对话伙伴，能在语言游戏中作出恰当的回应，那么这就会引发关于言语行为的前提条件，以及理解和意识在语言生成中所起作用的问题。哲学家和语言学家需要仔细重新审视他们对构成性概念的定义。在一个表现与能力越来越难以区分的时代，研究者需要思考如何定义“理解”；在这个时代，能否将理解归于高性能机器也已成为一个

备受关注的哲学争论话题（Butlin，2021年；Li 等人，2021年；Andreas，2022年；Frankish，2022年；Mitchell 和 Krakauer，2023年；Sobieszek 和 Price，2022年）。例如，我们的研究结果或许有助于促成关于语言输出生成的多重实现假说。

从推测的角度来看，对错误输出的分析有可能促进我们对人类认知能力的理解，这类似于通过研究视觉错觉来揭示视觉机制。人类会犯某些类型的错误，而人工系统会犯不同类型的错误，这说明了哪些潜在机制？由于这一研究领域发展迅速，现阶段很难得出明确的结论（Belinda 等，2021年）。然而，有许多迹象表明，仅基于 Transformer 架构的模型（即仅进行统计评估）无法可靠地取代人类认知能力。后续模型倾向于转而采用混合方法。不过，未来性能更优的模型或许能够凭借更通用、更具“经验主义”色彩的结构，以及比我们通常假定在人类认知能力背后存在的更少的“先天”或专门化架构，在复杂任务中取得成功。与此相关的是，我们可能很快就会面临这样一个问题：在有限领域内，意识和具身性是否确实是理解所必需的。

关于微调语言模型的版权法规定，是一个尚未有定论的伦理问题（例如，参见英国政府咨询，2021年）。未经作者许可，使用单一作者的作品对语言模型进行微调，是否属于对知识产权的合理使用，目前尚不明确。由于微调后的模型不太可能生成与作者语料库中的某段文本完全一致的长篇文字，通过微调语言模型借用观点，可能无法被识别为剽窃，即便这确实应被视为剽窃。因此，至少在法律问题明确之前，我们建议在使用某位作者受版权保护的文本进行微调或发布任何输出内容之前，先征得作者的明确许可。如何处理已故作者的作品，也应纳入考量（Nakagawa 和 Orita，2022年）。一种可能的做法是通过法律强制要求标注大语言模型的输出，以遏制学术欺诈、宣传和垃圾信息等滥用行为（例如，当前的《人工智能法案》草案——一项拟议中的欧盟法律（欧盟委员会，2021年）——要求对任何可能被误认为是人类互动的内容进行标注）。

过度依赖模型也是一种风险。尽管 DigiDan 的表现超出了我们的预期，但它并不能稳定地产生反映丹尼特观点的输出。这并不令人意外，因为深度学习

网络往往存在可靠性问题（Bosio 等人，2019年；Alshemali 和 Kalita，2020年）。在某些情况下，尽管输出模仿了丹尼特的风格，其内容却与丹尼特会认同的任何观点都大相径庭。缺乏经验的用户，或对目标作者作品不够熟悉的用户，可能会误以为，基于某位作者的作品微调的大语言模型，其输出很可能反映作者的真实观点，或作者会说的话（Bender 等人，2021年；Wedinger 等人，2021年）。对于学生、社交媒体用户，或其他宁愿向作者的微调模型提问也不愿阅读作者作品的人来说，这种想法尤其容易诱人。因此，我们建议，在向公众发布任何基于某位作者进行微调的语言模型之前都应谨慎行事，即使已获得作者许可也不例外。如果发布任何语言模型，就应明确说明其性质，指出其局限性，并明确标注所有输出均来自计算机程序，而非真人。如果将机器生成文本作为现有人物观点的引述或转述呈现，这可以说构成了伪造（丹尼特接受 Cukier 采访时所述，2022年；Strasser，2023年）。

随着机器生成文本越来越难以与人工生成文本区分开来，其他社会问题也会随之出现。未来教师如何确保提交的论文并非仅仅由语言模型生成（Herman，2022年；Hutson，2022年；Marche，2022年）？在聊天对话中，我们如何知道自己是在与人类而非聊天机器人互动？新的社会惯例或许会以证明某人确实是所写内容原作者为目标。大学或许会重新采用有监考的现场论文写作。机器生成内容越难与人类表达区分，滥用的风险就越大。例如，机器生成文本可能会在错误信息的传播中发挥重大作用（Marcus，2022年）。

除了大语言模型的进一步发展可能带来的巨大社会影响之外，此类模型的不可靠性也引发了关于认知结构的耐人寻味的问题。尽管通过扩大模型规模可以取得令人印象深刻的改进，但迄今为止的证据表明，完全由神经网络生成的输出仍将是不可靠的（Mitchell 和 Krakauer，2023年）。然而，尽管存在错误，通用模型仍可能具备与人类能力或通用智能大体相似的能力。这一问题促成了基准测试研究（Michael 等，2022；Talmor 等人，2020年；Webb 等人，2022年），也促成了大量关于如何揭示此类模型局限性的演示（Dou 等人，2022；

Marcus 和 Davies，2020年，2022年）。大型语言模型的输出在多大程度上、以及在何种条件下值得人们信任，这一问题很可能仍会存在争议。

在注意到这些 cautions 的同时，我们看到了微调大型语言模型的显著长期潜力。如果技术持续进步，采用混合技术的微调语言模型或许很快就能生成足够有趣的输出，成为专家进行精选的重要来源。可以将其与那些以特定作曲家的风格创作音乐的计算机程序（Hadjeres 等，2017；Daly，2021；Elgammal，2021）以及 Midjourney 等图像生成程序相比较。尽管这些输出中有许多并不出色，但经过筛选的输出可能具有相当高的音乐或艺术价值。作曲家或艺术家可以生成许多输出，选出最有潜力的作品，稍加编辑后，将其作为原创作品呈现——丹尼特本人曾在私人交流中提出过这种可能性。在这种情况下，语言模型将成为由人类使用的思维工具。同样，在哲学领域，专家可以使用特定的语料库对语言模型进行微调（例如，他们自己的语料库、某位喜爱的对话者或历史人物的语料库，或由多位精选哲学家的语料库汇编而成的集合），在各种提示词下生成多种输出，然后选出其中最有趣的内容，作为潜在思想的来源。

我们很难断言下棋机器对国际象棋有所信念。语言模型对哲学家抱有哲学信念的可能性则更低，尤其是当它们仍专注于下一个词预测，显然并没有关于世界的认知模型时。DigiDan 并不持有关于意识、上帝和动物痛苦的丹尼特式哲学观点。不过，没有哲学理解能力的机器或许能成为通往更高成就的跳板。也许我们正处于一个临界点，即将创造出能够生成文本的机器，这些文本看起来闪耀着哲学机敏、洞见或常识，可能会在读者心中激发新的哲学思想，也或许能为最终创造出真正具备哲学思维能力的人工实体铺平道路。

致谢
特别感谢丹尼尔·C·丹尼特和马修·克罗斯比。丹尼特在本项目的各个方面都给予了合作、建议和鼓励。马修·克罗斯比提供了专业技术支持，为本项目实施了微调，并与我们合作撰写了一篇为本项目奠定基础的概念性论文（Strasser、Crosby 和 Schwitzgebel, 2023年）。

参考文献

Alshemali, B., & Kalita, J.（2020年）。提升自然语言处理领域深度神经网络可靠性的综述。基于知识的系统，191，105210。doi:10.1016/j.knosys.2019.105210 Andreas, J.（2022年）。语言模型作为智能体模型。doi:10.48550/arXiv.2212.01681 Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., ... Shetty, S.（2019年）。利用低剂量胸部计算机断层扫描和三维深度学习进行端到端肺癌筛查。《自然医学》，25(6), 954–961。doi:10.1038/s41591-019- 0447-x Assael, Y., Shillingford, B., Whiteson, S., & Freitas, N.（2016年）。LipNet：句子级唇读。doi:10.48550/arXiv.1611.01599 Araoz, M.（2021年）。通过 GPT-3 采访阿尔伯特·爱因斯坦。https://maraoz.com/2021/03/14/einstein-gpt3/ Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S.（2021年）。随机鹦鹉的危险：语言模型会不会太大？FAccT '21：2021年 ACM公平、问责与透明度会议论文集， 610–623。doi:10.1145/3442188.3445922 Berwick, R.C., Pietroski, P., Yankama, B., & Chomsky, N.（2011年）。重新审视刺激贫乏。认知科学，*35*(7), 1207–1242。doi:10.1111/j.1551-6709.2011.01189.x Bosio, A., Bernardi, P., Ruospo, & Sanchez, E.（2019年）。深度神经网络的可靠性分析。2019年 IEEE拉丁美洲测试研讨会（LATS），1–6。doi:10.1109/LATW.2019.8704548 Brown, N., & Sandholm, T.（2019年）。多人扑克中的超人类人工智能。《科学》，*365*(6456)。doi:10.1126/science.aay2400 Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... Amodei, D.（2020年）。语言模型是少样本学习者。神经信息处理系统进展，33，1877–1901。doi:10.48550/arXiv.2005.14165 Brownlee, J.（2019年）。避免神经网络过度训练的提前停止简介。机器学习精通。
h
t
t
p
s://machinelearningmastery.com/early-stopping-to-avoid-overtraining-neural-network-models/

Buckner, C., & Garson, J.（2019年）。联结主义。载于 E.N. Zalta（编），《斯坦福哲学百科全书》。https://plato.stanford.edu/archives/fall2019/entries/connectionism Butlin, P.（2021年）。与机器分享我们的概念。Erkenntnis。doi:10.1007/s10670-021-00491-w Campbell, M., Hoane Jr, A. J., & Hsu, F. H.（2002年）。深蓝。人工智能，134(1–2)

，

57–83. Chomsky, N.（2007）。生物语言学探索：设计、发展、演化。《国际哲学研究杂志》，*15*(1), 1–21. Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., & Smith, N. A.（2021）。并非所有“人类的”都是黄金：评估对生成文本的人工评估。doi:10.48550/arXiv.2107.00061 Clarke, D.（2022）。Chat GPT-3：用它自己的话来说。https://www.verdict.co.uk/chat-gpt-3- interview/ Cukier, K.（2022）。Babbage：人工智能会拥有感知能力吗？《经济学人》。https://shows.acast.com/theeconomistbabbage/episodes/babbage-could-artificial-intelligence-become-sentientDaly, R.（2021）。人工智能软件创作新的 Nirvana 和 Amy Winehouse 歌曲，以提高人们对心理健康支持的认识。NME。 https://www.nme.com/news/music/ai-software-writes-new-nirvana-amy-winehouse-songs-raise-awareness-mental-health-support-2913524 Davidson, D.（1984）。《真理与解释探究》。牛津大学出版社。Descartes, R.（1649/1991）。致莫尔的信，1649年2月5日。载于 J. Cottingham, R. Stoothoff, D. Murdoch, and A. Kenny（编），《笛卡尔哲学著作》第3卷。剑桥大学出版社。Dou, Y., Forbes, M., Koncel-Kedziorski, R., Smith, N., & Yejin, C.（2022）。GPT-3生成的文本能否与人类文本区分开来？Scarecrow：一种审视机器文本的框架。计算语言学协会第60届年会论文集，7250–7274。doi:10.18653/v1/2022.acl-long.501。Dugan, L., Ippolito, D., Kirubarajan, A., & Callison-Burch, C.（2020）。RoFT：一种评估人类检测机器生成文本能力的工具。doi:10.48550/arXiv.2010.03070。

Elgammal, A.（2021）。音乐学家和计算机科学家团队如何完成了贝多芬未完成的第十交响曲。The Conversation。
https://theconversation.com/how-a-team-of-musicologists-and-computer-scientists-completed-beethovens-unfinished-10th-symphony-168160

Epstein, R., Roberts, G., & Beber, G.（2009）。《解析图灵测试》。Springer。doi:10.1007/978-1-4020-6710-5

欧盟委员会（21.4.2021）。人工智能法案。关于欧洲议会和理事会制定人工智能协调规则（《人工智能法案》）并修订若干欧盟立法法案的法规提案。https://artificialintelligenceact.eu/the-act/

Fawzi, A., Balog, M., Romera-Paredes, B., Hassabis, D., & Kohli, P.（2022年）。借助 AlphaTensor 发现新算法。https://www.deepmind.com/blog/discovering-novel-algorithms-with-alphatensor

Frankish, K.（2022年）。关于大语言模型的一些想法。《心智的把戏》博客文章。https://www.keithfrankish.com/blog/some-thoughts-on-llms/

Gao, C., Howard, F., Markov, N., Dyer, E., Ramesh, S., Luo, Y., & Pearson, A.(2022年)。使用人工智能生成内容检测器、抄袭检测器和盲法人工审阅者，对比 ChatGPT 生成的科学摘要与原始摘要。doi:10.1101/2022.12.23.521610

GitHub Copilot。https://docs.github.com/en/copilot

英国政府咨询（2021年）。人工智能意见征询：版权和相关权利。
https://www.gov.uk/government/consultations/artificial-intelligence-and-intellectual-property-call-for-views/artificial-intelligence-call-for-views-copyright-and-related-rights

Hadjeres, G., Pachet, F., & Nielsen, F. (2017)。DeepBach：一种可控的巴赫众赞歌生成模型。《第34届国际机器学习大会论文集》，1362–1371。

Harnad, S. (2003). 心智、机器与图灵。载于 J.H. Moor（编），《图灵测试》。认知系统研究（第253–273页）。Springer。doi:10.1007/978-94-010-0105-2_14

Herman, D.（2022年）。高中英语的终结。《大西洋月刊》。
https://www.theatlantic.com/technology/archive/2022/12/openai-chatgpt-writing-high-school-english-essay/672412/ Hilario, M.（1995年）。神经符号集成策略概述。会议论文集

e

联结主义—符号整合研讨会：从统一方法到混合方法。Hofstadter, D.（2022）。道格拉斯·霍夫施塔特认为，如今的人工神经网络并没有意识。《经济学人》。
https://www.economist.com/by-
i
n
v
itation/2022/06/09/artificial-neural-networks-today-are-not-conscious-according-to-douglas-hofstadter Huang, K.（2023）。人工智能聊天机器人引发警觉，大学开始全面改革教学方式。《纽约时报》。
https://www.nytimes.com/2023/01/16/technology/chatgpt-artificial-intelligence-universities.html Hutson, M.（2022）。人工智能能帮你写下一篇论文吗？《自然》，611(7934)，192–193。doi:10.1038/d41586-022-03479-w Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., ... Hassabis, D.（2021）。利用 AlphaFold 实现高精度蛋白质结构预测。《自然》，596(7873)，583–589。
doi:10.1038/s41586-021-03819-2 Klein, E.（2022）。这一刻比你想象的更加奇异。《纽约时报》。
https://www.nytimes.com/2022/06/19/opinion/its-not-the-future-we-cant-see.html
Li, B.Z., Nye, M., & Andreas, J.（2021）。神经语言模型中意义的隐式表征。计算语言学协会年会。Loebner, H.（2009）。如何举办图灵测试竞赛。载于 R. Epstein, G. Roberts, & G. Beber（编），《解析图灵测试》（第173–179页）。Springer。
doi:10.1007/978-1-4020-6710- 5 12_Mahian, O., Treutwein, M., Estellé, P., Wongwises, S., Wen, D., Lorenzini, G., ... Sahin, A.（2017）。学术语境中的相似性测量。《出版物》，5(3)，18，doi:10.3390/publications5030018 Mahdawi, A., & GPT-3（2020）。机器人写下了这整篇文章。人类，你现在害怕了吗？T

he

卫报。https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3

Mahowald, K., Ivanova, A., Blank, I., Kanwisher, N., Tenenbaum, J., & Fedorenko, E. (2023). 大型语言模型中的语言与思想分离：认知视角。doi:10.48550/arXiv.2301.06627 Marche, S. (2022). ChatGPT 会扼杀学生作文吗？《大西洋月刊》。
https://www.theatlantic.com/technology/archive/2022/12/chatgpt-ai-writing-college-student-essays/672371/Marcus, G. (2022). 像 ChatGPT 这样的人工智能平台易于使用，但也可能存在危险。《科学美国人》。
https://www.scientificamerican.com/article/ai-platforms-like-chatgpt-are-easy-to-use-but-also-potentially-dangerousMarcus, G., & Davis, E. (2020). GPT-3，夸夸其谈者：OpenAI 的语言生成器根本不知道自己在说什么。《麻省理工科技评论》。Marcus, G., & Davis, E. (2022). 像 ChatGPT 这样的大型语言模型，总会说出令人咋舌的话。
https://garymarcus.substack.com/p/large-language-models-like-chatgptMichael, J., Holtzman, A., Parrish, A., Mueller, A., Wang, A., Chen, A., ... & Bowman, S. R. (2022). 自然语言处理研究人员相信什么？自然语言处理社区元调查的结果。doi:10.48550/arXiv.2208.12852 Mitchell, M., & Krakauer, D. (2023). 关于人工智能大型语言模型的理解问题之争's。《美国国家科学院院刊》，*120*(13), e2215907120.doi:10.1073/pnas.2215907120Nakagawa, H., & Orita, A. (2022). 使用逝者个人数据。《人工智能与社会》，1–19

Novikova, J., Dušek, O. Curry, A.C., & Rieser, V. (2017年). 为什么自然语言生成需要新的评估指标。 2017年自然语言处理实证方法会议论文集。罗伯茨, M. (2022年). 谷歌的 LaMDA'是否是有感知能力的人工智能？这是个错误的问题。《华盛顿邮报》。
https://www.washingtonpost.com/opinions/2022/06/14/google-lamda-artificial-intelligence-sentient-wrong-question/Shanahan, M. (2023年). 谈论大型语言模型。doi:10.48550/arXiv.2212.03551 Schmidt, D., & Heckendorf, C. (2015). ngram软件包指南：快速 n元语法分词。

R

软件包。https://cran.r-project.org/web/packages/ngram/vignettes/ngram-guide.pdf Schneider, S.（2019年）。《人工的你》。普林斯顿大学出版社。

Schwitzgebel, E.（2021a）。越来越多的人或许很快会认为机器人具有意识并应享有权利。发表于博客 The Splintered Mind。https://schwitzsplinters.blogspot.com/2021/03/more-people-might-soon-think-robots-are.html Schwitzgebel, E.（2021b）。两篇由机器人生成的 The Splintered Mind 博文。发表于博客 The Splintered Mind。https://schwitzsplinters.blogspot.com/2021/11/two-robot-generated-splintered-mind.html Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., ... Hassabis, D.（2016年）。运用深度神经网络和树搜索掌握围棋。《自然》，529(7587)，484–489。doi:10.1038/nature16961 Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., van den Driessche, G., ... Hassabis, D.（2018年）。一种通过自我对弈掌握国际象棋、日本将棋和围棋的通用强化学习算法。《科学》，362(6419)，1140–1144。doi:10.1126/science.aar6404 Sobieszek, A., & Price, T.（2022年）。与人工智能玩游戏：GPT-3 及类似大型语言模型的局限。《心灵与机器》32，341–364。doi:10.1007/s11023-022- 09602-0 Steven, J., & Iziev, N.（2022年）。人工智能正在掌握语言。我们应该相信它所说的话吗？《纽约时报》。https://www.nytimes.com/2022/04/15/magazine/ai-language.html Strasser, A.（2022年）。从工具使用到社会互动。载于 J. Loh & W. Loh（编），《社会机器人学与美好生活》（第77-102页）。Transcript Verlag。doi:10.1515/9783839462652- 004 Strasser, A.（2023年）。复杂大型语言模型的陷阱（及优势）。doi:10.48550/arXiv.2303.1751 Strasser, A., Crosby, M., & Schwitzgebel, E.（2023年）。我们在创建一位哲学家的数字复制体方面能走多远？载于 R. Hakli, P. Mäkelä, & J. Seibt（编），《社会制度中的社会机器人》。机器人哲学2022论文集（第371–380页）。IOS Press。Talmor, A., Elazar, Y., Goldberg, Y., & Berant, J.（2020年）。oLMpics——语言模型预训练捕捉了什么。《计算语言学协会汇刊》，8，743–758。Turing, A. M.（1950年）。计算机器与智能。《心灵》，59（236），433–460。https://doi.org/10.1093/mind/LIX.236.433

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H. T., ... Le, Q.（2022年）。LaMDA——用于对话应用的语言模型。doi:10.48550/arXiv.2201.08239

Tiku, T.（2022年）。这位谷歌工程师认为，公司的人工智能已经有了生命。《华盛顿邮报》。https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/

van der Lee, C., Gatt, A., van Miltenburg, E., Wubben, S., & Krahmer, E. (2019年). 自动生成文本人工评估的最佳实践。第十二届国际自然语言生成会议论文集，355–368, 东京，日本。计算语言学协会。Webb, T., Holyoak, K., & Lu, H. (2022年). 大型语言模型中涌现的类比推理。doi:10.48550/arXiv.2212.09196 Wiseman, H. (2020年). 哲学家戴维·查尔默斯接受采访，探讨新型人工智能文本生成器 GPT3 是否可能具有意识。Facebook 帖文。https://www.facebook.com/howard.wiseman.9/posts/4489589021058960Zimmerman, A. (编) (2020 年).《哲学家论 GPT-3 》（附 GPT-3 回复更新版）。发表于博客 *Daily Nous*。https://dailynous.com/2020/07/30/philosophers-gpt-3

表1

GPT-3 生成结果与丹尼特训练语料库之间匹配的六个或更多单词的字符串。“出现次数”列表示该短语在训练语料库中出现于多少个不同的训练数据片段中。较短字符串的出现次数总计不包括较长匹配字符串中的出现次数。（因此，如果表中出现了某个较大 n 元语法的子集，那么这意味着它在文本中独立出现过，而非仅出现在较大的 n 元语法中。例如，“思考的直觉泵及其他工具”在“我的新书《思考的直觉泵及其他工具》中”之外还出现过一次。）

字符串	词数	出现次数
在我的新书《直觉泵与其他思考工具》中	11	1
其组织方式使它	8	1
并不存在这样的东西	7	10
弄清楚它应该做什么	7	1
直觉泵和其他思考工具	7	1
并不存在这样的东西	6	14
我学到了很多	6	2
以这样的方式组织	6	2
从经验中学习的能力	6	1
但如果你想获得	6	1
从经验中学习的能力，我们	6	1
在我的书《打破魔咒》中	6	1
以这样的方式使它	6	1
其组织方式是	6	1
我的书《打破魔咒》，我	6	1
当然，这就引出了一个问题	6	1
也就是说，存在	6	1
并非显而易见的是	6	1
留给……的空间就越大	6	1
陷入……的陷阱	6	1
鉴于……应当做什么	6	1