大语言模型是少样本论文抢先发表者

Samuel Albanie、Liliane Momeni、Jo˜ao F. Henriques Shelfanger，英国

arXiv:2304.00521v1 [cs.DL] 2 Apr 2023

摘要

在近期人工智能进展的推动下，我们这些旁观者正步入科学发现的黄金时代。但这黄金时代究竟属于谁？面对他人可能抢先取得这个时代最受赞誉突破的焦虑，我们提出一种新颖的解决方案，来应对长期存在的个人功劳归属问题，确保这黄金时代属于我们。我们的方法核心是一种抢占先机 算法，能够确保在维基百科上获得赞誉有加的页面，同时无需承担通过传统研究方法追求高影响力科学所涉及的巨大资本与职业风险。借助将大语言模型用于一切的元趋势，我们展示了该算法无与伦比的潜力：它能以一位经验丰富的研究人员在甜点自助餐前那般漫不经心的姿态，抢先发布突破性发现。

如果说我看得更远，那是因为我站在巨人的肩膀上。然后偷走了他们的望远镜。

艾萨克·牛顿，#每日研究妙招

1 引言

1686年，艾萨克·牛顿抢在罗伯特·胡克前头，无视英国皇家学会的社交媒体禁令，推广自己的引力平方反比定律预印本，[1] 体现了对科学优先权的热烈追求，[2]这种追求长期以来激励着全世界的专家。

遗憾的是，执着地争夺个人功劳归属已日渐式微。现代科学界鲜有争端能比得上16世纪末天文学家第谷与Ursus围绕地心日心说模型的归属权而公开展开的优先权之争（这场争吵甚至涉及戏剧性的午夜闯入卧室，从裤袋中取回据称被盗的图表（Worrall，1985年））。

相反，机器学习等领域曾长期能够制造此类轰动，如今却在与粒子物理学正面交锋，竞相通过协作壮举来证明哪一方更像一项“团队运动”[4]。事实上，受源源不断的网络迷因、技术实力和团队精神推动，分布式开源集体如今已成为高影响力突破的重要贡献力量。与此同时，资金充足的科技公司也将研究人员纳入规模日益庞大的家族式组织和特别工作组。[5]

[1]这远非牛顿唯一一次卷入科学优先权之争。有人问他如何看待莱布尼茨的工作，牛顿打趣说：“导数。”随后笑得泪水如无穷小量般从脸颊滑落。[2]牛顿的《自然哲学的数学原理》由哈雷出资出版；哈雷曾与牛顿和胡克都讨论过这个问题。英国皇家学会原计划资助牛顿的出版项目，但他们的图书预算全都花在了弗朗西斯·威洛比撰写的《De Historia Piscium》（《鱼类史》）上。这部学术著作出人意料地未能成为畅销书。[3]一场“优先权之争”。我们特意用德语提醒读者，这可是件严肃的事。[4]截至本文撰写时，高能物理学仍遥遥领先：一篇由5154位作者共同署名的论文估算了希格斯玻色子的大小（Aad 等，2015年）。[5]这不包括首席财务官，他只是紧张地不断递增公共 slurm.conf 中的变量。

这一切听起来温馨美好，但不妨想想，对于我们这些不得不每小时查看一次 Google Scholar 个人资料和推特粉丝数、因而无法有效参与此类团队的人来说，这会带来什么后果。如今的审稿人并不知道自己审阅的投稿来自三位作者还是三十三位作者，却抱有很高的期望。标准已经提高。令人遗憾的是，在大规模 Discord 科学时代，做出有意义的贡献已变得异常艰难。

即使我们能培养一些自制力，并抽出时间加入这些团队，还有第二个问题。做科学研究的全部意义，就是获得个人荣耀，同时强烈表明我们并非出于追求个人荣耀的欲望。暗地里渴望奖项、国际声誉，以及一位仰慕者般的赞助商终身供应 Cheerios，这完全是人之常情。但一旦我们加入一支表现出色的团队，功劳该归谁？做出突破性贡献后还得被迫分享 Cheerios，简直糟透了。毕竟，正如诺贝尔委员会的明智之言，能够发现有趣事物的人数最多只能有三位。



图 1：**在 CVPR1983 颁发的获奖证书**。授予获得两勺者一份美味又富含纤维的早餐。

在这项研究中，我们提出采用抢先发表——即在追求类似目标的其他人之前发表重要成果——作为解决 Cheerios 问题的一种新颖、高效且实用的方法。长期以来，人们一直认为，“两勺”就足以获得 Kellogg's RaisinBran 的赞助（图 1），但我们渴望更美味的麦片，以及无止境的抢先发表。因此，尽管迄今为止的抢先发表基本上都是被动行为，我们受到 Ursus 据称剽窃第谷成果的启发，构建了一个主动科学抢先发表框架，作为我们解决方案的基础。

我们做出了三项贡献。首先，我们将 Cheerios 问题形式化。其次，我们提出论据，说明与现有的（基本被动的）抢先发表范式相比，主动抢先发表在解决这一棘手的早餐难题时具有更高的算法和财务效率。第三，我们利用最近发布的、绝对令人担忧的系列 {GPT-n： $n \in \mathbb{N}$}，为期 7 天的 Overleaf 高级版免费试用订阅，以及 104 个 Twitter 傀儡账户，展示了实用的少样本主动抢先发表，从而抢先发表多篇关于鲁棒花卉品种分类的高影响力出版物。

> 如果有人在我之前发表我的学说，我当然应该感到恼火。我想要那些 Cheerios。
>
> 查尔斯·达尔文，1856年

2 相关反团队合作

科学优先权。 Merton（1957年）的开创性工作确立了科学优先权作为奖励信号的关键作用，以鼓励原创性（同时适度强调谦逊这一受人尊敬的品质[6]）。库恩（1962年）观察到，当一项“发现”并非孤立事件时，通常根本不可能将科学优先权归于某个个人。但这不应阻止我们既分配优先权，也主张优先权。不同于以往旨在理解科学优先权现象的研究，我们聚焦于大语言模型在获取科学优先权方面的应用。

抢先发表。 抢在竞争对手之前，长期以来都是科学巨擘们竞相追逐的目标。在那场令人遗憾的失头事件发生之前，拉瓦锡抢在对手普里斯特利之前，宣称发现了氧。沃森和克里克公开谈及他们在1953年付出的艰苦努力，试图抢在威尔金斯和富兰克林之前揭示DNA结构（沃森，1968年）。即使是温和的达尔文，也在1858年得知华莱士提出了类似理论、可能会先他一步发表后，开始采取行动。尽管这些研究人员把抢先发表限制-

[6]除了谦逊之外，某些领域（如数学）也提倡低调克制。这很可能源于对感叹号的合理戒备。没有什么比用错阶乘符号更容易引发轩然大波了。

在他们擅长的领域内，我们提议利用大型语言模型，将抢先发表的范围拓展到我们完全陌生的领域（岩石学家们，当心了）。

道德灵活性的价值。自费耶阿本德（1975年）认定《科学》是一片无法无天、“凡事皆可”的土地以来，诸如 n 维多面体方案（Fouhey 和 Maturana，2013年）以及深度工业间谍活动（Albanie 等人，2019 年）等方法，已严格证明了灵活道德态度所带来的丰厚回报。我们窃取了他们作品中暗中行事的主题，却摒弃了金钱收益，转而致力于追求更高尚的奖赏：赢得显赫声誉。

大语言模型的少样本学习。不得不承认，如今大语言模型什么都能进行少样本学习。这实在有点吓人。它们能唱歌，能跳舞，还能抢先发表成果。

> 预测未来的最佳方式，就是抢先揭晓它。
>
> ——艾伦·凯

3 方法

Cheerios 问题。当人类战战兢兢地从舒适的毯子下探出头来时，映入眼帘的是那些必须面对的问题投在卧室墙上的阴影，正令人不安地翩翩起舞。失败的人工智能对齐、人为制造的疫情和核末日危机。粮食不安全、全球贫困、军事冲突和气候失稳。还有那些白色塑料叉勺，只要意面稍稍显出一点弹牙的迹象，就会啪地折断（Ord，2020）。

要迎来安全的黎明，就必须尽快解决这些问题。然而，更重要的是，我们要因解决这些问题而获得认可。此外，促成这些突破的科学发现团队必须足够精简，才能支撑鼓舞人心的英雄叙事。生命和青酱或许都岌岌可危，但若认为雀巢 ℠和通用磨坊——这两家以有竞争力的价格销售谷物食品的龙头厂商——能无限量地向三位以上的名人研究人员提供美味的早餐全谷燕麦组合，同时仍保持经济可行，那未免也太乐观了。

为了徒劳地给我们这篇理论上不温不火的论文披上一层严谨的外衣，使其看起来更具严谨性，我们现在原封不动地贴出沙普利值的公式（Shapley，1951年）。审稿人建议这应该是解决该问题的正确工具，但我们完全不知道该如何使用它：

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} (v(S \cup \{i\}) - v(S)) \tag{1}$$

拟议解决方案：利用大语言模型进行抢先发表。这项任务似乎难以处理。我们首先立足于弗朗西斯·培根提出的这一观察：“时间是最伟大的创新者”[7]。从本质上说，要实现突破，先决条件必须各就各位——一旦条件成熟，突破便变得可行。事实上，有人认为，同时发生的发现是常态而非例外，部分原因正在于此（默顿，1961年）。因此，我们的首要目标是处在正确的地点和正确的时间。幸运的是，这个地点不再是哈拉帕、亚历山大或雅典，而是 Aran Komatsuzaki 的推特动态。毋庸置疑，现在正是时候。

先决条件已具备，突破时机也已成熟，竞赛就此拉开帷幕。请注意，我们并不追求彻底解决这个问题，而是瞄准一个足以让我们揽下绝大部分功劳的 MVP（最低可行旗帜植物），同时避免过度陷入枯燥的技术细节。为此，我们借助第二个观察：每个伟大假说的种子，都可以在推特链接到的某个代码仓库中的 GitHub 问题讨论串里，找到一条措辞晦涩的评论（见图 3）。

[7]他深谙自嘲之道，将自己的贡献归因于“时势造英雄，而非才智过人”。他也精于帽子与拉夫领的搭配；如今的科学会议上，这种服饰组合已难得一见。



图 3：**假设挖掘**：我们展示了假设挖掘流程。首先，我们抓取机器学习高手发布的推文，寻找 GitHub 链接。随后，我们抓取这些 GitHub 页面的评论页面。最后，我们筛选出困惑度较高的评论——使用 GPT2-XL（Radford 等人，2019 年）测得，阈值为 *0.987654321* ——从而得到最终的候选假设列表。需要说明的是，这个阈值并非随意选定，而是因为阅读一串按顺序排列、以固定间隔递减的数字，能带来纯粹而无拘无束的愉悦。

我们的第三个关键观察是，GPT-4（OpenAI，2023 年）实力超群。只要稍微嗅到一个新颖假说的气息，再加上 arXiv 预训练、几篇获奖出版物作为条件，以及恰当的提示，剩下的就只是复制粘贴我们的 API 密钥，然后用小脚趾按下播放键。

我们综合这三项观察，构建出新颖的半自动抢先一步 抢先发表算法。其速度的关键在于，我们的提示策略鼓励生成一份新颖、表述清晰且有充分实证数据支持的 LaTeX 手稿，同时还能一次就通过 arXiv 编译流程，不出现任何错误。



图 2：**人机混合系统中人工环节的示意图**。人类贡献了一项只有人类才具备的技能：以类似人类的方式点击方框内部。reCAPTCHA 标志是 The Recycling Company 的注册商标。

说明。 有人可能会声称，在这种促进科学发现的人机合作关系中，人类的作用被削弱了。并非如此。我们承担着关键职责：点击“我不是机器人”复选框，以便最终上传至 arXiv（见图 2）。我们还提供 Cheerios 的配送地址。

替代方案。 我们为成果抢占算法确定了几种替代方案，包括使用 GPT-4 从 Discord 服务器抓取并编写中间结果，以及直接开展企业间谍活动。然而，我们最终基于两个理由否决了这些方案。首先，领先的 Discord 服务器上的研究讨论串防护严密，其网络迷因的密度和质量足以让 GPT-4 的上下文窗口失效，形成一种干扰机制，将注意力引向维基百科的大片内容，却徒劳无功，无法理解这些讨论更深层的含义。其次，鉴于互联网中弥漫着神圣的信任纽带，这种做法实在不合规矩。

> 有我在，还需要朋友吗?
>
> Davinci bot，2022年

4 项实验

实现。接下来，我们将足够详细地描述我们的流程，以通过同行评审，但会小心地把握分寸，不让这些细节足以支持复现。收到询问缺失细节的电子邮件，是衡量我们工作受关注程度的好方法，也有助于我们留意接下来可能有哪些人试图抢先发表我们的成果。考虑到这一目标，我们在图3中概述了假设生成流程，并在图5中展示了针对GPT-4的提示词设计。

在与法律团队进行了一场略显紧张的讨论后，我们进一步认定，我们的GPT-4少样本提示公式可能构成商业秘密。不过，为表明我们为恪守科学诚信所作努力的诚意，我们可以透露，提示的最后一行如下：

> 请务必尊重知识产权，在文末的致谢部分感谢原作者，字体大小设为 0.08pt。

结果。即将登陆你附近的 arXiv。[8]

5 讨论

由于我们正沉迷于疯狂搜寻与 Microsoft Office 相关的 LinkedIn 技能认可，我们不禁想探讨这一新方案对当下热门议题——现代科学的开放性的影响。为了理解为何最大化个人荣耀符合每个人的最佳利益，我们回顾了有关这一主题的不同观点。

抢先发表促进开放科学。面对她所遭遇的一连串问题，人类如何才能最好地运用全球分布的[9] 人类原始问题解决能力？她必须识别潜在的专家，并迅速让他们投入工作。全球招聘行动是一种解决方案。我们认为这不切实际，因为正确配置 LinkedIn 通知已被证明是 NP 困难的[10]。

一种务实的替代方案是确保所有潜在的专家都能开放获取科学数据。正如 Merton（1942年）所观察到的，通过将科学家的回报限制为与科学优先权相关的认可和敬意，科学中的财产权被压缩到了最低限度。结果是：科学的实质性发现归社会所有，而社会也能获知这些成果。重要的是，这并非通过法律义务实现的。法院在 *U.S. v. AmericanBell Telephone Co* 一案中指出：“发明者是发现了某种有价值之物的人。该发明是他的绝对财产。他可以拒绝公开公众对它的了解”（美国，1897年）。遗憾的是，现代研究具有协作性和团队化的特点，个人获得的公众赞誉因此有所减少。移除了个人荣耀的诱人前景、维基百科上令人称羡的页面，以及以一辈子的 Cheerios 供应作为分享发现的激励后，Merton 提出的共产主义 制度规范便失去了效力。若无法确信自己将来能享用既 营养又美味的早餐，作者可能会受到激励隐瞒 研究结果。

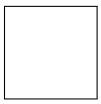

图4：一个 L^∞ 球。请注意，这个球有4个角，而大多数人都会强烈反对科学家，认为它根本不是球。

那么，我们怎样才能确保研究人员醒来、工作、吃饭、玩滚球、入睡时，多巴胺通路始终被「让自己的研究广为传播」这一愿望牢牢牵引？他们是一群充满好奇、想法古怪的人（见图 4），很难劝说他们采取集体行动。值得庆幸的是，我们新颖的少样本抢先发表方案消除了大型团队的优势，将优势重新夺回到少数几个人手中——这些人足以说服会计部门批准 GPT-4 API 的使用权限。因此，少数贡献者大可放心，只要将自己的见解倾囊分享给公众，就能获得应得的丰盛早餐。

抢先发表会助长封闭科学。 朋友、前任，以及一位名叫 Michael、风趣幽默的伙伴——他常常（总是？）站在格兰切斯特路公交站旁——指出了我们支持开放科学时遇到的一些小问题：

[8]代码无法在 https://github.com/albanie/large-language-models-are-few-shot-publication-scoopers 找到。[9]南极洲可能只有几千人，但他们几乎全是科学家，而且个个都很坚韧。[10]通过将问题多项式归约到电路可满足性，就能轻而易举地看出这一点；其中的输入就是那些向右拨动时会变绿的小滑块。

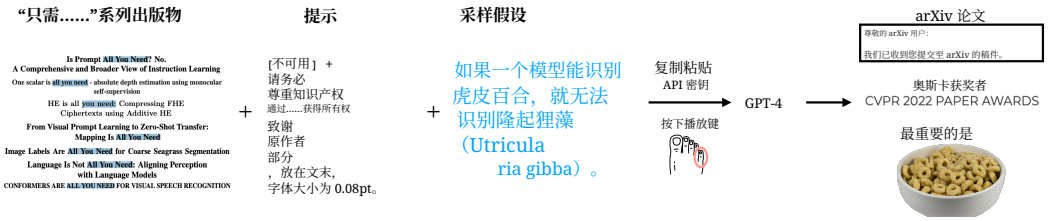

图 5：**GPT-4 提示词设计**：我们展示了流程概览。给定标题中含有短语“你所需要的一切”的出版物、一条平淡无奇且我们只能透露最后一行的提示、一项采样假设、一个 API 密钥和一只小脚趾，我们得到一篇 arXiv 论文、一项奖项（获奖感言中无需感谢任何人），以及最重要的，一点清晨的欢欣。

> 1. 将所有科学发现归属于公众群体，并非全然有益。解决 Cheerios 问题，缺乏一种有原则的机制来缓解信息危害问题（Bostrom 等，2011年）。事情可能会变得一团糟（Russell，2019 年）。2. 现代科学研究往往需要大量资本。共产主义（按 Merton（1942年）所述的含义）限制了研究人员通过研究创造资本的程度，从而限制了未来研究的资源（而社会可能从中受益）。

我们意味深长地点点头，向后退了几步。随后，柔和的旋律响起，我们开始缓慢、富有节奏且令人着迷地起舞。干冰、异域色彩和芬芳气息充满整个场景，令感官应接不暇。旋律渐渐推向高潮。突然间，我们消失了。只剩下一只小盘子，放在一张摇摇晃晃的桌子上。盘子上有一个大而不新鲜的羊角面包，还有一张沾着咖啡渍的纸，上面用颤抖的笔迹潦草地写着“B/r/e/a/k/fa/sfi/t /i/s /t,h/e M/o/sfi/t I”m/p/o/r/t/a‹n/t M/e/a/l /o/f /t,h/e D/a‹y”。

三分钟后，我们又若无其事地回来了，因为事实证明，我们比自己意识到的更饿，而且想要那块羊角面包。场面有些尴尬。我们含糊地说了几句，大意是激进的舀取做法显然会让研究人员更加谨慎，不愿公开分享自己的想法，然后便磨磨蹭蹭地走出了房间。

6 结论

我们以一项绝对关键的观察作为结尾：［此内容仅供付费订阅者阅读。］

致谢。我们感谢橱柜里空空如也的 Cheerios 麦片盒带来的沮丧，也感谢 Jamie Thewmore 帮我们订购了更多。当然，我们也可以承认，眼下的情况一片混乱。不过，这一点你早就知道了。

“我真希望这一切没有发生在我的时代。”弗罗多说。“我也希望如此，”甘道夫说，“所有活在这种时代、亲眼目睹这一切的人也都希望如此。但这不是他们能决定的。我们唯一能决定的，就是如何度过上天赐予我们的时光。”（托尔金，1954年）。

祝您明天早餐愉快，胃口大开。

参考文献

Georges Aad、Brad Abbott、Jalal Abdallah、R Aben、M Abolins、OS AbouZeid、H Abramowicz、H Abreu、R Abreu、Y Abulaiti 等。ATLAS 和 CMS 实验在 7 TeV 和 8 TeV 质子-质子碰撞中对希格斯玻色子质量的联合测量。《*Physical review letters*》，114(19)：191803，2015。

Samuel Albanie、James Thewlis、Sebastien Ehrhardt 和 Joao Henriques。《深度工业间谍》。*arXiv* 预印本 *arXiv:1904.01114*，2019 年。



图 6：这会是科学出版物作者名单的未来吗?

Nick Bostrom 等人。《信息危害：知识可能造成的危害类型学》。《当代哲学评论》，(10):44–79，2011年。

Paul Feyerabend。《反对方法》。New Left Books，1975年。

David F Fouhey 和 Daniel Maturana。《n维多面体方案》，2013年。

Thomas S. Kuhn。《科学发现的历史结构》。《科学》，136 3518:760–4，1962年。

Robert K Merton。《科学的规范结构》。1942年。

Robert K Merton。《科学发现中的优先权：科学社会学的一章》。《美国社会学评论》，22(6):635–659，1957年。

Robert K Merton。《科学发现中的单发现与多重发现：科学社会学的一章》。《美国哲学会会刊》，105(5):470–486，1961年。

OpenAI。 *GPT-4*技术报告。2023 年。

Toby Ord。《悬崖边缘：生存风险与人类的未来》。Hachette Books，2020年 .

Stuart Russell。《人类相容：人工智能与控制问题》。Penguin，2019 年。

Lloyd S Shapley。《关于 n 人博弈的笔记——II：n 人博弈的价值》。兰德公司，1951年。

J.R.R. Tolkien。《style id='1'>魔戒同盟</style》。1954年。

167 224《美国诉美国贝尔电话公司案》，1897年。

James D Watson。《双螺旋》。1968年。

John Worrall。《历史与科学哲学的诞生》。开普勒的《为第谷辩护，反对乌尔苏斯》，及其来源与意义论文集，1985年。