GPTs是GPTs。对大型语言模型的劳动力市场影响潜力的早期观察

Tyna Eloundou[1], Sam Manning[1,2], Pamela Mishkin[1], and Daniel Rock[3]

1OpenAI
2OpenResearch
3宾夕法尼亚大学

2023年3月21日

摘要

我们调查了生成性预训练转化器（GPT）模型和相关技术对美国劳动力市场的潜在影响。使用一个新的评分标准，我们根据职业与GPT能力的对应关系进行评估，其中包括人类的专业知识和GPT-4的分类。我们的研究结果表明，大约80%的美国劳动力可能至少有10%的工作任务受到GPT的影响，而大约19%的工人可能看到他们至少50%的任务受到影响。这种影响跨越了所有的工资水平，高收入的工作可能会面临更大的影响。值得注意的是，这种影响并不限于近期生产力增长较高的行业。我们的结论是，生成性预训练变压器表现出通用技术（GPTs）的特征，这表明这些模型可能具有明显的经济、社会和政策影响。

1 简介

如图1所示，最近几年、几个月和几周，生成性人工智能和大型语言模型（LLMs）领域取得了显著进展。虽然公众经常将LLMs与生成式预训练转化器（GPT）的各种迭代联系起来，但LLMs可以使用一系列架构进行训练，并不限于基于转化器的模型（Devlin等人，2019）。LLM可以处理和产生各种形式的序列数据，包括汇编语言、蛋白质序列和国际象棋游戏，不仅仅是自然语言的应用。在本文中，我们在某种程度上交替使用LLM和GPT，并在我们的评分标准中明确指出，这些应被视为类似于通过ChatGPT或OpenAI Playground提供的GPT-家族模型（在标注时包括GPT-3.5家族的模型，但不包括GPT-4家族）。我们研究具有文本和代码生成能力的GPT，并采用 "生成性人工智能 "一词来额外包括图像或音频等模式。

不过，我们的研究与其说是出于这些模型的进展，不如说是出于我们在围绕它们开发的补充技术中看到的广度、规模和能力。补充技术的作用还有待观察，但最大化LLM的影响似乎取决于将

它们与更大的系统整合（Bresnahan，2019；Agrawal等人，2021）。虽然我们将讨论的重点放在LLM的生成能力上，但通过将LLM用于其他任务，可能会有新型的软件和机器通信，包括像嵌入这样的事情，使

*通讯作者（pamela@openai.com）。作者贡献相同，按字母顺序排列。

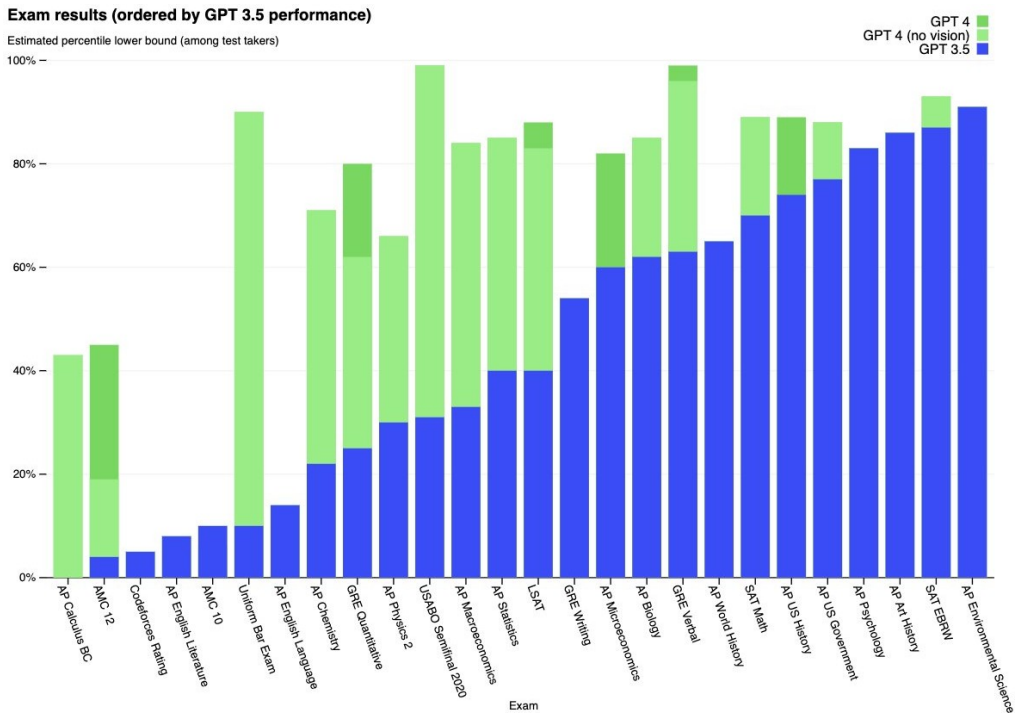


图1：为了了解模型能力的进展速度--考虑GPT-3.5和GPT-4之间考试成绩的跳跃性（OpenAI, 2023b）。

它可以建立自定义的搜索应用或像总结和分类这样的任务，在这些任务中，不清楚什么是或不是生成性的区别。

为了使这一进展的背景和补充技术的劳动影响预测，我们提出了一个新的评分标准，以了解LLM能力及其对工作的潜在影响。这个评分标准（A.1）衡量了任务对GPT的整体暴露，遵循了之前量化机器学习暴露的工作精神（Brynjolfsson等人，2018；Felten等人，2018；Webb，2020）。我们将暴露度定义为潜在经济影响的代理，而不区分劳动力增加或劳动力转移的影响。我们采用人类注释者和GPT-4本身作为分类器，将这一标准应用于美国经济中的职业数据，主要来自O*NET数据库[12]。

为了构建我们的主要曝光数据集，我们同时收集了人类注释和GPT-4分类，使用了一个经过调整的提示，以便与作者的标签样本达成一致。我们观察到，当汇总到任务层面时，GPT-4的反应以及人类和机器的评价之间有类似的一致水平。这一措施反映了对使人类劳动更有效率的技术能力的估计；然而，社会、经济、监管或其他决定因素意味着技术可行性并不能保证劳动生产率或自动化结果。我们的分析表明，当考虑到目前的模型能力和在此基础上建立的预期工具时，大约19%的工作至少有50%的任务被暴露。人工评估表明，当考虑到现有的语言和代码能力而没有额外的软件或模式时，只有3%的美国工人有超过一半的任务暴露在GPT中。考虑到其他生成模型和补充技术，我们的人类评估表明

[1] 这与最近利用先进的语言模型来模拟人类行为的社会科学研究不同 （Horton，2023；Sorensen等，2022）。

[2]虽然我们的接触标准不一定将语言模型的概念与任何特定的模型联系在一起，但我们观察到的GPT-4的能力以及我们看到的与OpenAI的发射伙伴一起开发的一套能力强烈地激励着我们（OpenAI, 2023b）。

高达49%的工人可能有一半或更多的任务接触到LLMs。

我们的研究结果一致表明，在人类和GPT-4注释中，大多数职业都在一定程度上表现出对LLM的暴露，不同类型的工作暴露程度不同。工资较高的职业一般都有较高的暴露度，这一结果与对整体机器学习暴露度的类似评估相反（Brynjolfsson等人，2023）。当使用O*NET的技能表对暴露度进行回归时，我们发现，严重依赖科学和批判性思维技能的角色与暴露度呈负相关，而编程和写作技能与LLM暴露度呈正相关。继Autor等人（2022a）之后，我们研究了 "工作区 "的进入障碍，并发现职业接触LLM的机会随着工作准备的难度增加而微弱。换句话说，在工作中面临较高（较低）进入壁垒的工人往往会更多（更少）接触到LLMs。我们进一步将我们的测量方法与以前记录经济中自动化风险分布的工作进行比较，发现结果大致一致。我们研究的大多数其他技术暴露措施与我们首选的暴露措施有统计学上的显著相关性，而人工常规性和机器人技术暴露的措施则显示出负相关。这些早期的努力（Acemoglu和Autor，2011a；Frey和Osborne，2017；Brynjolfsson等人，2018；Felten等人，2018；Webb，2020；Brynjolfsson等人，2023），连同工资控制，所解释的方差从60%到72%不等，表明我们的AI暴露措施的28%至40%的变化仍然没有被以前的技术暴露所说明测量。

我们按行业分析风险，发现信息处理行业（4位数NAICS）表现出高风险，而制造业、农业和采矿业表现出较低风险。过去十年的生产力增长和整体GPT暴露之间的联系似乎很弱，这表明一个潜在的乐观的情况，即未来LLM的生产力增长可能不会加剧可能的成本疾病效应（Baumol，2012）。[3]

我们的分析表明，像GPT-4这样的LLMs的影响可能是普遍存在的。虽然LLM的能力随着时间的推移不断提高，但即使我们今天停止开发新的能力，其不断增长的经济效应预计也会持续存在并增加。我们还发现，当我们考虑到补充性技术的发展时，LLM的潜在影响会显著扩大。总的来说，这些特征意味着生成性预训练变换器（GPTs）是通用技术（GPTs）4（Bresnahan和Trajtenberg，1995；Lipsey等人，2005）。(Goldfarb等人，2023）认为，机器学习作为一个广泛的类别，很可能是一种通用的技术。我们的证据支持更广泛的影响，因为即使是机器学习软件的子集也能独立满足通用技术地位的标准。本文的主要贡献是提供了一套衡量LLM影响潜力的方法，并展示了应用LLM来有效地、大规模地开发这种测量方法的使用案例。此外，我们还展示了LLM的通用潜力。如果 "通用技术就是通用技术"，那么对于政策制定者来说，LLM发展和应用的最终轨迹可能是难以预测和监管的。与其他通用技术一样，这些算法的大部分潜力将在广泛的有经济价值的用例中出现，包括创造新的工作类型（Acemoglu和Restrepo，2018；Autor等人，2022a）我们的研究有助于衡量现在技术上的可行性，但必然会错过LLMs随着时间推移不断发展的影响潜力。

本文的结构如下。第2节回顾了先前的相关工作，第3节讨论了方法和数据收集，第4节介绍了汇总统计和结果，第5节将我们的测量与先前的工作联系起来，第6节探讨了结果，第7节提供了结论意见。

[3] 鲍莫尔的成本病是一种理论，它解释了为什么劳动密集型服务，如医疗保健和教育的成本会随着时间的推移而增加

。发生这种情况是因为其他行业的技术工人的工资增加了，但这些服务行业的生产力或效率却没有相应的提高。因此，这些行业的劳动成本与经济中的其他商品和服务相比，变得相对 。

4在本文的其余部分，我们用GPT来泛指大型语言模型，如通过OpenAI提供的那些模型为例，当它被用于说明 "GPT是GPT "之外时，我们拼出通用技术。

2 文献回顾

2.1 大型语言模型的进步

近年来，大型语言模型（LLMs）在人工智能（AI）研究领域崭露头角，展示了其处理各种复杂语言任务的能力。这一进展是由多种因素推动的，包括增加模型参数数、更大的训练数据量和增强的训练配置（Brown等人，2020；Radford等人，2019；Hernandez等人，2021；Kaplan等人，2020）。广泛的、最先进的LLMs，如LaMDA（Thoppilan等人，2022）和GPT-4（OpenAI，2023b），在翻译、分类、创意写作和代码生成等不同应用中表现出色--这些能力以前需要由专家工程师使用特定领域的数据开发专门的、特定任务的模型。

同时，研究人员利用微调和带有人类反馈的强化学习等方法改善了这些模型的可引导性、可靠性和实用性（Ouyang等人，2022；Bai等人，2022）。这些进步增强了模型辨别用户意图的能力，使其更加方便用户和实用。此外，最近的研究揭示了LLMs编程和控制其他数字工具的潜力，如API、搜索引擎，甚至其他生成性人工智能系统（Schick等人，2023；Mialon等人，2023；Chase，2022）。这使得各个组件的无缝整合能够获得更好的效用、性能和概括性。从长远来看，这些趋势表明，LLM可能有能力执行任何通常在计算机上执行的任务。

在大多数情况下，生成性人工智能模型主要被部署为模块专家，执行特定的任务，如从标题生成图像或从语音转录文本。然而，我们认为，必须采用更广泛的视角，将LLM视为额外工具的关键构建块。虽然构建这些工具并将其整合到综合系统中需要时间，并需要对整个经济领域的现有流程进行重大的重新配置，但我们已经观察到正在出现的采用趋势。尽管有其局限性，法律硕士正越来越多地被整合到写作协助、编码和法律研究等领域的专门应用中，为企业和个人更广泛地采用GPT铺平了道路。

我们强调这些补充技术的意义，部分原因是，由于事实不准确、固有偏见、隐私问题和虚假信息风险等问题，开箱即用的通用GPT对于各种任务可能仍然不可靠（Abid等人，2021；Schramowski等人，2022；Goldstein等人，2023；OpenAI，2023a）。然而，专门的工作流程--包括工具、软件或人在环形系统--可以通过纳入特定领域的专业知识来帮助解决这些缺陷。例如，Casetext提供基于LLM的法律研究工具，为律师提供更快、更准确的法律研究结果，利用嵌入和总结来对抗GPT-4提供关于法律案件或文件集的不准确细节的风险。GitHub Copilot是一个编码助手，它采用LLM来生成代码片段和自动完成代码，然后用户可以根据他们的专业知识接受或拒绝。换句话说，虽然GPT-4本身确实不 "知道现在是什么时候"，但给它看一下也很容易。

此外，当LLM超过一个特定的性能阈值时，可能会出现一个积极的反馈循环，使他们能够协助建立工具，以提高其在各种情况下的有用性和可用性。这可以降低创建此类工具所需的成本和工程专业知识，有可能进一步加速LLM的采用和整合。(Chen等人，2021年；Peng等人，2023年）LLM也可以成为机器学习模型开发中的宝贵资产--作为研究人员的编码助手、数据标签服务或合成数据生成器。这些模型有可能为任务层面的经济决策做出贡献，例如，通过完善人类和机器之间

的任务和子任务分配方法（Singla等人，2015；Shahaf和Horvitz，2010）。随着LLMs随着时间的推移不断改进，并更好地与用户的偏好保持一致，我们可以预见性能会不断增强。

然而，必须认识到，这些趋势也带来了各种严重的风险。(Khlaaf等人, 2022; Weidinger等人, 2022; Solaiman等人, 2019)

2.2 自动化技术的经济影响

大量且不断增长的文献涉及广义的人工智能和自动化技术对劳动力市场的影响。以技能为导向的技术变革概念和自动化的任务模型--通常被认为是理解技术对劳动力影响的标准框架--起源于研究表明，技术进步提高了对熟练工人的需求，而不是非熟练工人（Katz和Murphy，1992）。许多研究都建立在这一概念之上，在基于任务的框架内探索技术变革和自动化对工人的影响（Autor等人，2003；Acemoglu和Autor，2011b；Acemoglu和Restrepo，2018）。这方面的研究表明，从事常规和重复性任务的工人在技术驱动的迁移中面临着更大的风险，这种现象被称为常规偏向的技术变革。最近的研究区分了技术的任务替代效应和任务恢复效应（新技术增加了对更多劳动密集型任务的需求）（Acemoglu and Restrepo, 2018, 2019）。一些研究表明，自动化技术导致了美国的工资不平等，由专门从事常规任务的工人的相对工资下降驱动（Autor等人，2006；Van Reenen，2011；Acemoglu和Restrepo，2022b）。

之前的研究采用了各种方法来估计人工智能能力与工人在不同职业中承担的任务和活动之间的重叠。这些方法包括将专利描述映射到工人的任务描述（Webb，2020；Meindl等人，2021），将人工智能能力与O*NET数据库中记录的职业能力联系起来（Felten等人，2018，2023），通过认知能力将人工智能任务基准评估与工人任务相一致（Tolan等人。2021），标记美国职业子集的自动化潜力，并使用机器学习分类器来估计所有其他美国职业的潜力（Frey和Osborne，2017），对任务级自动化进行建模，并将结果汇总到职业级见解（Arntz等人，2017），专家预测（Grace等人，2018），以及与本文最相关的，设计一个新的评分标准来评估工人活动对机器学习的适用性（Brynjolfsson等人，2018，2023）。其中一些方法发现，在任务层面上对人工智能技术的接触往往在职业中是多样化的。考虑到每项工作都是一捆任务，很少能找到人工智能工具可以完成几乎所有工作的职业。(Autor等人，2022a）也发现，自动化和增强的暴露往往是正相关的。还有越来越多的研究考察了LLM的具体经济影响和机会（Bommasani等人，2021；Felten等人，2023；Korinek，2023；Mollick和Mollick，2022；Noy和张，2023；彭等人，2023）。在开展这项工作的同时，我们的测量结果有助于描述语言模型与劳动力市场更广泛的潜在关联性。

通用技术（如印刷、蒸汽机）（GPTs）的特点是广泛扩散、持续改进和产生互补性创新（Bresnahan和Trajtenberg，1995；Lipsey等，2005）。它们的深远后果在几十年内展开，很难预测，特别是在劳动力需求方面（Bessen，2018；Korinek和Stiglitz，2018；Acemoglu等人，2020；Benzell等人，2021）。实现通用技术的全部潜力需要广泛的共同发明（Bresnahan和Trajtenberg，1995；Bresnahan等人，1996，2002；Lipsey等人，2005；Dixon等人，2021），这是一个昂贵和耗时的过程，涉及发现新的商业程序（大卫，1990；Bresnahan，1999；弗雷，2019；Brynjolfsson等人，2021；Feigenbaum和Gross，2021）。因此，许多关于机器学习技术的研究集中在系统层面的采

用，认为组织系统可能需要重新设计，以有效利用新的机器学习进展（Bresnahan，2019；Agrawal等人，2021；Goldfarb等人，2023）。适当设计的系统可以产生相当大的商业价值并提高公司业绩（Rock，2019；Babina等人，2021；Zolas等人，2021），人工智能工具促进了发现过程（Cockburn等人，2018；Cheng等人，2022）。通过

任务	编号职业名称	DWAs	任务描述
	14675计算机系统师，以	监测计算机系统性能的工程师/架构确保正常运行。	监测系统运行，以发现潜在的问题。
	18310急性护理护士	操作诊断性或治疗性疾病。医疗器械或设备。准备使用的医疗用品或设备。	设置、操作或监测侵入性设备和装置，如结肠造口或气管切开设备、机械呼吸机、导管、胃肠管和中央管。
4668.0	赌笼工人	执行销售或其他金融交易。	为顾客兑现支票和处理信用卡预付款。
	15709网上招商	执行销售或其他金融交易。	交付已完成的交易和装运的电子邮件
	教师，特殊教育除外	–	让家长志愿者和高年级学生参与儿童活动，以促进参与集中、复杂的游戏。
	6568小学教师，特殊教育除外	–	让家长志愿者和高年级学生参与儿童活动，以促进参与有重点的复杂游戏。

表1：O*NET数据库中的职业、任务和详细工作活动样本。我们看到，仅对活动进行汇总是不准确的，这一点可以从以下事实中得到证明：我们希望赌笼工人能够亲自完成给定的详细工作活动，并使用一些体力，而我们希望网上商人能够仅仅通过电脑完成同样的活动。

采用任务层面的信息来评估LLM是否符合GPT标准，我们试图将这两个角度合并起来，以理解技术与劳动的关系。

我们试图以几种方式在这些不同的文献流的基础上进行研究。与Felten等人（2023）一样，我们将分析重点放在LLM的影响上，而不是更广泛地讨论机器学习或自动化技术。此外，我们提出了一种新的方法，采用LLMs，特别是GPT-4，来评估任务的暴露和自动化潜力，从而加强人类的评分工作。随后，我们将我们的发现汇总到职业和行业，捕捉到当代美国劳动力市场的整体潜在风险。

3 方法和数据收集

3.1 美国各职业所从事的活动和任务数据

我们使用O*NET 27.2数据库（O*NET，2023），其中包含了1016个职业的信息，包括其各自的详细工作活动（DWA）和任务。DWA是一个综合行动，是完成任务的一部分，如 "研究脚本以确定项目要求"。另一方面，任务是一个特定职业的工作单位，可能与任何、一个或多个DWA相关。我们在表1中提供了一个任务和DWAs的样本。我们使用的两个数据集包括。

- 19,265项任务，其中每项任务都有一个 "任务描述 "和一个相应的职业，大多数任务都与一个或多个DWA相关。
- 2,087个DWAs，其中大多数DWAs与一个或多个任务相连，任务可能与一个或多个DWAs相关，但有些任务缺乏任何相关的DWAs。

3.2 工资、就业和人口统计学方面的数据

我们从劳工统计局提供的2020年和2021年的职业就业系列中获得就业和工资数据。这个数据集包含了职业名称，每个职业的工人数量，以及每个职业的工资。

职业，以及2031年的职业级就业预测，进入一个职业所需的典型教育和达到一个职业能力所需的在职培训（BLS，2022）。我们使用BLS推荐的与O*NET的对照表（BLS，2023b）来连接O*NET任务和DWA数据集和BLS劳动力人口统计（BLS，2023a），后者来自于当前人口调查（CPS）。这两个数据源都是由美国政府收集的，主要是捕捉那些非自雇的、有证件的、在所谓正规经济中工作的工人。

3.3 曝光

我们根据暴露度评分标准来展示我们的结果，其中我们将**暴露度定义**为衡量访问GPT或GPT驱动的系统是否会将人类执行特定DWA或完成任务所需的时间减少至少50%。我们在下面提供了一个评分标准的摘要，而完整的评分标准可以在A.1中找到。当我们有DWAs的标签时，我们首先在任务层面进行汇总，然后再在职业层面进行汇总。

曝光概述

没有接触（E0），如果。
- 在保持同等质量的情况下，完成活动或任务所需的时间没有减少或减少到最低程度，或
- 根据以下标准，使用所述能力的任何组合都会降低活动/任务产出的质量。

直接接触（E1），如果。
- 仅仅使用理论上的LLM或通过ChatGPT或OpenAI操场描述的GPT-4，可以将完成DWA或任务所需的时间减少至少一半（50%）。

LLM+曝光（E2），如果。
- 仅仅获得法学硕士学位并不能将完成活动/任务所需的时间至少减少一半，但
- 可以在LLM上开发更多的软件，可以将高质量完成特定活动/任务的时间至少减少一半。在这些系统中，我们算作访问图像生成系统。

[a] 在实践中，从A.1中的完整评分表可以看出，我们将图像能力的获取单独分类（E3），以方便注释，尽管我们将E2和E3结合起来进行所有分析。

我们将曝光阈值设定为：在保持质量稳定的前提下，完成一项具体的工作任务所需的时间可能减少50%。我们预计，对于那些实现了相当大的生产力增长的应用来说，采用率将是最高和最直接的。虽然这个阈值有些随意，但我们选择它是为了便于注释者解释。

然后，我们收集了人类和GPT-4生成的注释，这些注释是本文中大部分分析的基础。

- *人类评分。*我们通过对每个O*NET详细工人活动（DWA）和所有O*NET任务的一个子集应用评分标准来获得人类注释，然后将这些DWA和任务汇总起来。

r 此外，无论选择什么样的阈值，我们猜测现实世界中任务时间的减少可能会略低于或大大低于我们的估计，这导致

我们选择了一个相对较高的阈值。在我们自己的验证标签中，我们发现这与GPT或由GPT驱动的应用程序是否能执行任务的核心部分或几乎整个任务密切相关。

比较	□	加权	协议	pearson's
GPT-4, Rubric 1; 人类	□	E1	80.8%	0.223
	□	E1 + .5*E2	65.6%	0.591
	□	E1 + E2	82.1%	0.654
GPT-4, Rubric 2; 人类	□	E1	81.8%	0.221
	□	E1 + .5*E2	65.6%	0.538
	□	E1 + E2	79.5%	0.589
GPT-4, Rubric 1; GPT-4, Rubric 2	□	E1	91.1%	0.611
	□	E1 + .5*E2	76.0%	0.705
	□	E1 + E2	82.4%	0.680

表2：模型和人类的一致性和皮尔逊相关分数的比较。一致性分数是通过观察两组人对注释的一致程度来确定的（例如E0、E1或E2）。在本文中，我们使用GPT-4，Rubric 1。

在任务和职业层面上的得分9。为了确保这些注释的质量，作者亲自对大量的任务和DWA进行了标注，并招募了经验丰富的人类注释者，他们作为OpenAI对标工作的一部分，对GPT输出进行了广泛的审查（Ouyang等人，2022）。

- *GPT-4评分。*我们对早期版本的GPT-4（OpenAI，2023b）进行了类似的评分，但对所有的任务/职业对而不是DWAs进行了评分。我们对评分标准（在这种情况下被用作对模型的 "提示"）做了轻微的修改，以加强与一组人类标签的一致性。表2中给出了完整的一致率。

我们为我们感兴趣的因变量构建了三个主要的衡量标准：(i) □，对应于上述暴露标准中的E1，预计代表一个职业中暴露任务比例的下限，' (ii) □，是E1和0.5*E2之和，其中E2的0.(iii) ζ，是E1和E2之和，是暴露的上限，提供了对GPT和GPT驱动的软件的最大暴露的评估。我们在表2中总结了注释组和措施之间的一致性。在其余的分析中，如果没有特别说明，读者可以认为我们指的是β暴露--这意味着所有通过ChatGPT或OpenAI Playground等工具直接暴露的任务被认为是需要一些补充性创新的任务的两倍。

3.4 我们方法的局限性

3.4.1 人类的主观判断

我们的方法的一个基本限制在于标签的主观性。在我们的研究中，我们采用了熟悉GPT模型能力的注释者。然而，这个群体的职业并不多样化，可能会导致对GPT的可靠性和在不熟悉的职业中执行任务的有效性作出有偏见的判断。我们认识到，要为一个职业中的每项任务获得高质量的标签，需要从事这些职业的工人，或者至少要拥有深入的知识。

9 作者对明显需要高度体力或手工灵活性的DWA进行了注释，签约的注释者对其余的活动进行了标注，还有一个任务子集，包括那些没有相关DWA的任务和那些在汇总DWA注释后没有 明确任务级注释的任务。

'尽管最近多模态GPT模型取得了进展（OpenAI，2023b），但视觉能力并没有包括在评估范围内。

□接触。

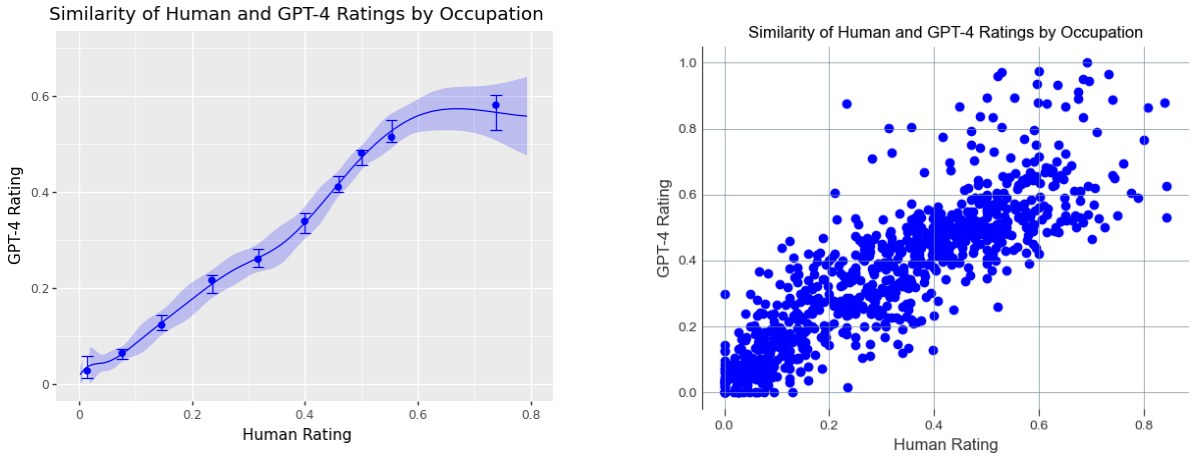


图2：人类评分者（X轴）和GPT-4评分（Y轴）显示出对职业的GPT暴露有高度的一致性。在按照□方法将暴露分数汇总到职业的最高水平附近，GPT-4评分往往低于人类评分。我们列出了原始散点和二进制散点。在接近暴露等级的高端时，人类平均更有可能将一个职业评为暴露等级。

这些职业中的不同任务。这代表了未来验证这些结果的一个重要工作领域。

3.4.2 用GPT-4测量GPTs

最近的研究表明，GPT-4作为一个有效的判别器，能够应用复杂的分类法，并对措辞和重点的变化做出反应。(OpenAI, 2023b)GPT-4任务分类的结果对评分标准的措辞、提示的顺序和组成、评分标准中是否有具体的例子、提供的详细程度以及关键术语的定义等方面的改变很敏感。根据小型验证集的观察结果，对提示进行迭代，可以提高模型输出和评分标准意图之间的一致性。因此，呈现给人类的评分标准和用于GPT-4的评分标准之间存在轻微的差异。这一决定是有意做出的，以引导模型走向合理的标签，而不过度影响人类注释者。因此，我们使用了多个注释来源，但相对于其他来源，没有一个应该被认为是最终的基础真理。在分析中，我们将把来自人类注释者的结果作为我们的主要结果。在为LLM分类制定有效的评分标准方面，进一步的改进和创新仍然是可能的。尽管如此，我们仍然观察到人类评分和GPT-4评分在职业层面上关于GPT系统的总体接触的高度一致（见表2，图2）。

3.4.3 额外的弱点

- **基于任务的框架的有效性。**目前还不清楚职业在多大程度上可以被完全分解为任务，也不清楚这种方法是否系统地遗漏了某些类别的技能或任务，而这些技能或任务是胜任一项工作所默示的。此外，任务可以由子任务组成，其中一些任务的自动化程度比其他任务高。有些任务可以作为工作的前奏。

  其他任务，因此，下游任务的完成取决于前驱任务。如果基于任务的细分确实不能有效代表

一个职业中大多数工作的执行方式，那么我们的暴露分析就会在很大程度上失效。

- **相对与绝对措施。**最好将这些措施解释为相对措施，例如，一个估计为0.6的职业可能应被解释为只是比0.1的职业暴露得多。

- **缺少专业知识和任务解释。**在标注过程中，人类注释者大多不知道映射到每个DWA的具体职业。这导致了任务和职业的聚合逻辑不明确，以及标签中一些明显的差异，见表1。我们尝试了各种聚合方法，发现即使使用最大匹配度的
  如果存在匹配的人类<>模型标签，那么协议仍然相对一致。最终，我们为存在重大分歧的任务/职业对收集了额外的标签。

- **具有前瞻性，并会发生变化，有一些早期证据。**准确预测未来的LLM应用仍然是一个巨大的挑战，即使对专家来说也是如此（OpenAI，2023b）。新出现的能力、人类感知的偏差和技术发展的转变都会影响准确性
  关于LLM对工人任务的潜在影响的预测的可靠性。我们的预测本质上是前瞻性的，是基于当前的趋势、证据和对技术可能性的看法。因此，它们可能会随着该领域的新进展而改变。例如，一些今天看来不太可能对LLM产生影响的任务可能会随着新模型能力的引入而改变。反之，那些看起来已经暴露的任务可能会面临不可预见的挑战，限制语言模型的应用。

- **分歧的来源。**虽然我们没有严格检查分歧的来源，但我们发现有几个地方，人类和模型在评估中往往会 "卡壳"。

  – 一些任务或活动，虽然理论上法律硕士可以帮助或完成任务，但采用法律硕士来完成这些任务需要多人改变他们的习惯或期望（如会议、谈判）。

  – 目前有一些需要人类监督的规定或暗示人类判断或同情的规范的任务或活动（例如，做决定、咨询），以及

  – 已经存在可以合理地将任务自动化的技术的任务或活动（例如，进行预订）。

4 结果

通用技术是比较少见的，其特点是普遍性，随着时间的推移而改进，并发展出重要的共同发明和外溢效应（Lipsey等人，2005）。我们对GPTs（生成性预训练变压器）对劳动力市场影响的评估是有限的，因为它没有考虑全要素生产率或资本投入潜力。除了对劳动力的影响外，GPTs还可能影响这些方面。

在这个阶段，某些GPT标准比其他标准更容易评估。例如，从长远来看，评估这些模型能力的长期影响以及互补性应用和系统的增长更为可行。我们在这个早期阶段的主要重点是测试GPT语言模型对经济有普遍影响的假设，类似于（Goldfarb等人，2023）通过招聘信息对机器学习扩散的分析，以评估机器学习作为一个算法类别的GPT潜力。与其使用招聘信息或研究一般的机器学习，不如用人类和GPT注释来研究任务评估方法，可能会发现GPT的影响是否只限于一小部分类似的

任务或职业。

我们的研究结果表明，基于其任务层面的能力，通用技术有可能对美国经济中的各种职业产生重大影响，显示了通用技术的一个关键属性。在下面的章节中，我们将讨论各种角色和工资结构的结果。关于美国经济中各行业相对暴露的其他结果可在附录D中找到。

4.1 统计摘要

这些措施的统计摘要可在表3中找到。人类和GPT-4注释都表明，职业层面的平均□值在0.14和0.15之间，表明对于中位职业，大约15%的任务直接暴露于GPTs。这个数字在□中增加到30%以上，在□中超过50%。巧合的是，人类和GPT-4注释也将数据集中总任务的15%至14%标记为暴露于GPTs。

根据□值，我们估计80%的工人属于至少有一项任务暴露于GPTs的职业，而19%的工人属于有一半以上的任务被标记为暴露的职业。尽管任务受影响的可能性很大，但GPT必须被纳入更广泛的系统，以充分实现这一潜力。正如通用技术所常见的那样，这种共同发明的障碍可能会阻碍GPTs在经济应用中的快速推广。此外，预测对人类监督的需求是具有挑战性的，特别是对于模型能力等于或超过人类水平的任务。虽然对人类监督的要求最初可能会减慢采用和扩散的速度，但GPT和GPT驱动的系统的用户可能会越来越熟悉该技术随着时间的推移，特别是在了解何时和如何信任其产出方面。

职业级别 暴露于人体				
			GPT-4	
平均值	中位数	平均数	中位数	
□	0.14	0.14	0.14	0.16
□	0.30	0.21	0.34	0.22
□	0.46	0.30	0.55	0.34
任务级别 暴露于人体				
			GPT-4	
平均值	中位数	平均数	中位数	
□	0.15	0.36	0.14	0.35
□	0.31	0.37	0.35	0.35
□	0.47	0.50	0.56	0.50

表3：我们的人类和模型暴露数据的汇总统计。

4.2 工资和就业

在图3中，我们展示了整个经济的暴露强度。第一个图显示的是工人总数的暴露，而第二个图显示的是职业总数的暴露。图中的每一点都代表了Y轴上的工人（和职业）的估计百分比，X轴上的暴露程度（□，□，和□）。例如，人类注释者确定，2.4%的工人是$_{50}$ -exposed□, 18.6%是□$_{50}$ -exposed, 49.6%是□$_{50}$ -exposed, 其中50%的阈值来源于

x轴，工人的百分比来自图2右图的y轴。在任何给定的点上
在x轴上，□和□之间的垂直距离代表了在直接接触GPT之外的工具和应用所带来的接触潜力。暴露的分布在工人和雇员中都是相似的。

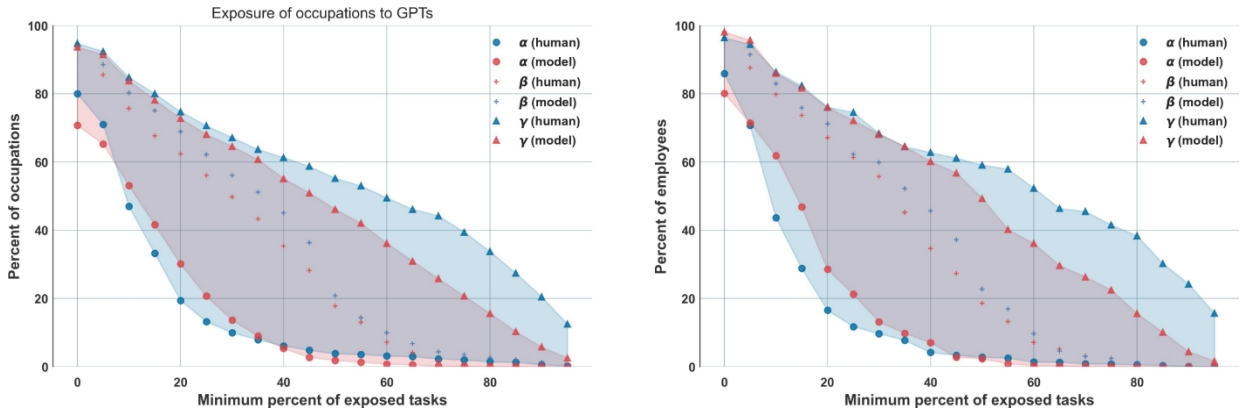


图3：整个经济体的暴露强度，左边显示为受影响职业的百分比，右边显示为受影响工人的百分比。不同职业和不同工人的暴露分布相似，这表明工人在职业中的集中度与职业对GPT或GPT驱动的软件的暴露并不高度相关。然而，我们确实预计它可能与为特定领域开发GPT驱动的软件的投资有更高的相关性。

职业，表明工人在职业上的集中度与职业上接触GPT或GPT驱动的软件没有很大的关联性。

如图4所示，在职业层面上，人类和GPT-4的注释表现出质量上的相似性，并趋于相关。与GPT-4注释相比，人类注释对高工资职业的暴露估计略低。虽然有许多低工资职业的暴露量高，而高工资职业的暴露量低，但二元散点图的总体趋势显示，工资越高，GPT暴露量越大。

对GPT的潜在暴露似乎与目前的就业水平没有什么关联。在图4中，人类和GPT-4对总体暴露的评价都被汇总到职业层面（Y轴），并与总就业人数的对数（X轴）进行比较。这两张图都没有显示出不同就业水平的GPT暴露的明显差异。

4.3 技能的重要性

在这一节中，我们研究了一项技能对一个职业的重要性（如O*NET数据集中的注释）和我们的暴露度量之间的关系。我们首先采用O*NET提供的基本技能（技能定义可在附录B中找到），并对每个职业的技能重要性进行归一化处理，以提高可解释性。然后，我们对我们的暴露措施（□，□，□）进行回归分析，检查技能重要性和暴露之间的关联强度。

我们的研究结果表明，**科学**和**批判性**思维能力的重要性与暴露度呈强烈的负相关，这表明需要这些技能的职业不太可能受到当前语言模式的影响。相反，**编程**和**写作**技能与接触率呈强烈的正相关，意味着涉及这些技能的职业更容易受到语言模式的影响（详细结果见表5）。

4.4 进入的障碍

接下来，我们研究了进入的障碍，以更好地了解是否存在因工作类型而产生的接触差异。一个这样的代理是O*NET的职业级别描述符，称为 "工作区"。一个工作区将

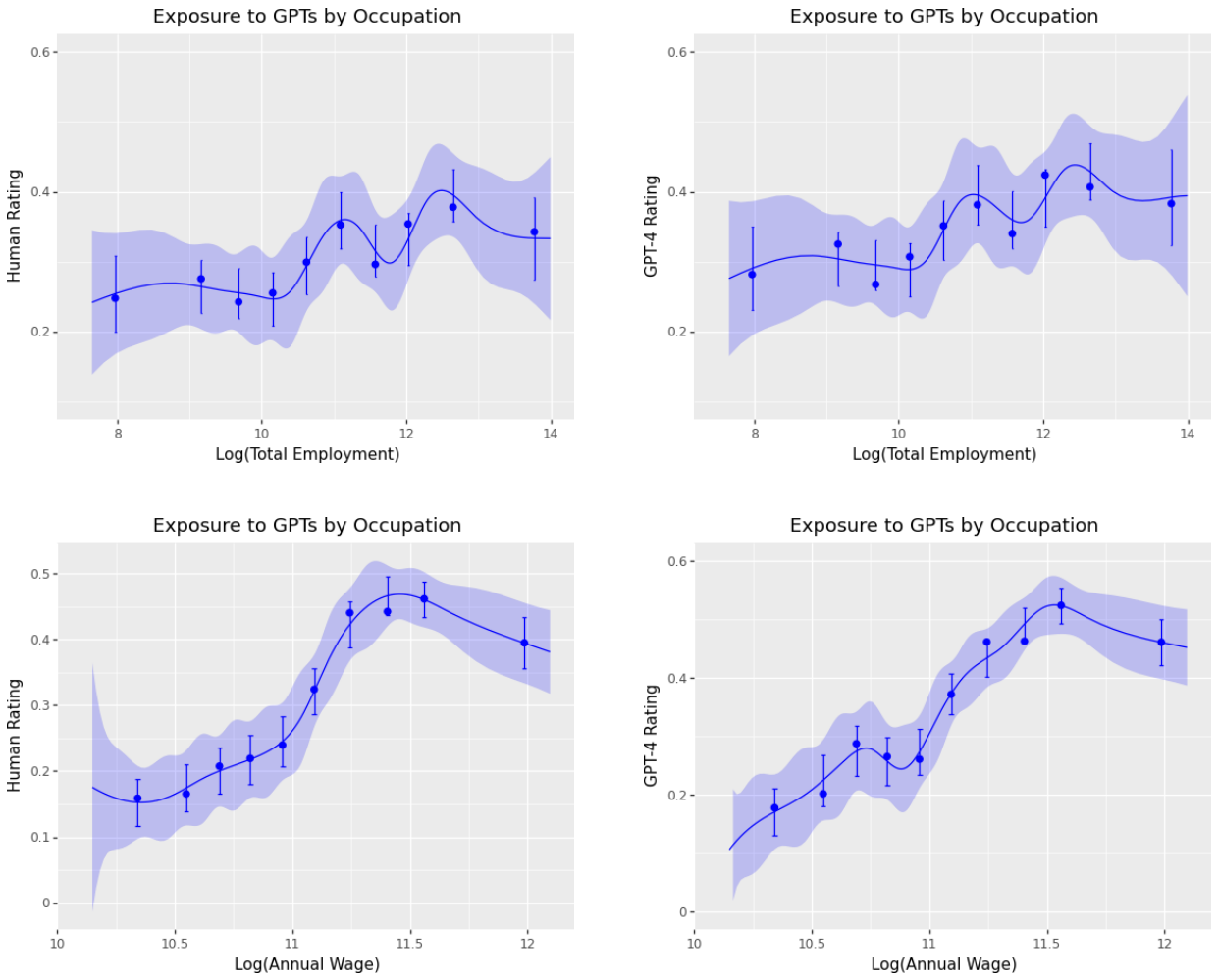


图4：二元散点图描述了由人类评估员和GPT-4评估的各种职业中的语言模型（LLMs）的接触情况。这些图将职业层面上的GPT暴露（□）与职业内总就业人数的对数和职业年薪中值的对数进行比较。虽然存在一些差异，但人类和GPT-4的评估都表明，工资较高的职业往往更容易受到LLM的影响。此外，根据我们的评分标准，许多工资较低的职业表现出较高的曝光率。在计算平均暴露分数时，核心任务的权重是职业内补充任务的两倍。就业和工资数据来自于2021年5月进行的BLS-OES调查。

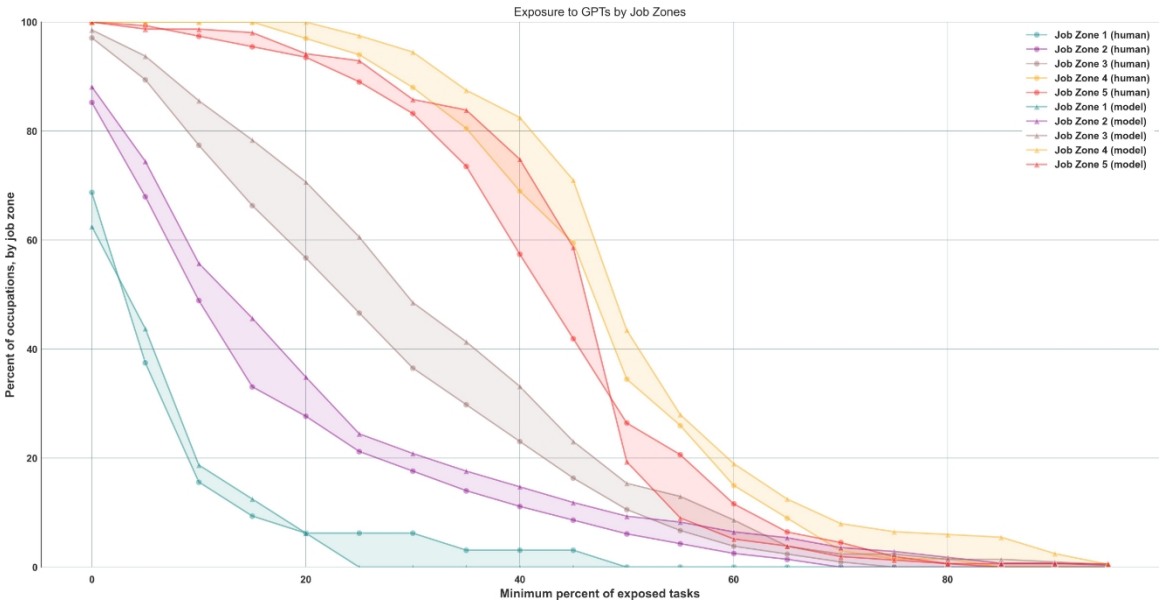


图5：五个工作区的职业暴露评级，这是一组类似的职业，根据从事这些职业所需的教育、经验和在职培训水平来分类。

在（a）获得该职业工作所需的教育水平，（b）从事该工作所需的相关经验数量，以及（c）从事该工作所需的在职培训程度方面具有相似性的职业。在ONET数据库中，有5个工作区，工作区1需要最少的准备工作（3个月），工作区5需要最广泛的准备工作，即4年或更长时间。我们观察到，随着所需准备程度的增加，各工作区的收入中位数呈单调增长，第1工作区的工人收入中位数为30，230美元，第5工作区的工人收入中位数为80，980美元。

我们所有的衡量标准（□、□和□）都显示了一个相同的模式，即从工作区1开始，暴露度增加。到工作区4，在工作区5保持相似或减少。与图3中的5类似，我们绘制了处于每个暴露阈值的工人的百分比。我们发现，平均来说，在工作区1到5，职业暴露超过50%的工人百分比□分别为0.00%（工作区1），6.11%（工作区2），10.57%（工作区3），34.5%（工作区4）和26.45%（工作区5）。

4.4.1 入职所需的典型教育

由于纳入就业区既考虑了所需的教育--这本身就是技能获取的替代物--也考虑了所需的准备，我们寻求数据来分解这些变量。我们使用劳工统计局的职业数据中的两个变量。"入职所需的典型教育"和 "达到能力所需的在职培训 "的职业。通过研究这些因素，我们旨在发现对劳动力有潜在影响的趋势。有3,504,000名工人，我们缺乏关于教育和在职培训要求的数据，因此，他们没有被列入汇总表。

我们的分析表明，拥有学士、硕士和专业学位的人比没有正式学历的人更容易接触到GPT和GPT驱动的软件（见表7）。有趣的是，我们还发现拥有一些大学教育但没有学位的人对GPT和

GPT驱动的软件表现出较高的接触水平。在检查显示进入壁垒的表格时，我们发现，接触最少的工作需要最长的培训，一旦达到能力，可能会提供较低的回报（以中位数收入计算）。相反，不需要在职培训或只需要实习/居留的工作似乎可以获得更高的收入，但更容易受到GPT的影响。

集团	接触最多的职业	曝光率百分比
人类 □	口译员和笔译员	76.5
	调查研究人员	75.0
	诗人、抒情诗人和创意作家	68.8
	动物科学家	66.7
	公共关系专家	66.7
人类 □	调查研究人员	84.4
	作家和作者	82.5
	口译员和笔译员	82.4
	公共关系专家	80.6
	动物科学家	77.8
人类 □	数学家	100.0
	报税人	100.0
	金融定量分析员	100.0
	作家和作者	100.0
	网页和数字界面设计者	100.0
	人类将15种职业标记为 "完全暴露"。	
型号 □	数学家	100.0
	函授书记员	95.2
	区块链工程师	94.1
	法庭报告员和同声传译员	92.9
	校对员和抄写员	90.9
型号 □	数学家	100.0
	区块链工程师	97.1
	法庭报告员和同声传译员	96.4
	校对员和抄写员	95.5
	函授书记员	95.2
型号 □	会计师和审计师	100.0
	新闻分析师、记者和新闻工作者	100.0
	法律秘书和行政助理	100.0
	临床数据管理员	100.0
	气候变化政策分析师	100.0
	该模型将86种职业标记为 "完全暴露"。	
最高的差异	搜索营销策略师	14.5
	平面设计师	13.4
	投资基金经理	13.0
	财务经理	13.0
	保险评估师，汽车损坏	12.6

表4：根据每个测量值，暴露程度最高的职业。最后一行列出了具有最高$\square^2$值的职业，表明它们在脆弱性预测方面具有最大的可变性。暴露百分比表示职业任务中暴露于GPT（α）或GPT驱动的软件（β和ζ）的份额，其中暴露被定义为推动完成任务的时间减少至少50%（见暴露评分表A.1.因此，本表所列的职业是那些我们估计GPT和GPT驱动的软件能够为工人节省大量时间来完成一项任务的职业。

他们的任务的很大一部分，但这并不一定表明他们的任务可以被这些技术完全自动化。

基本技能	□ (std err)	□ (std err)	□ (std err)
	所有的技能重要性得分都被归一化，在0和1之间。		
恒定	0.082*** (0.011)	-0.112*** (0.011)	0.300*** (0.057)
积极倾听	0.128** (0.047)	0.214*** (0.043)	0.449*** (0.027)
数学	-0.127*** (0.026)	0.161*** (0.021)	0.787*** (0.049)
阅读理解	0.153*** (0.041)	0.470*** (0.037)	-0.346*** (0.017)
科学	-0.114*** (0.014)	-0.230*** (0.012)	-0.346*** (0.017)
讲话	-0.028 (0.039)	0.133*** (0.033)	0.294*** (0.042)
写作	0.368*** (0.042)	0.467*** (0.037)	0.566*** (0.047)
主动学习	-0.157*** (0.027)	-0.065** (0.024)	0.028 (0.032)
批判性思维	-0.264*** (0.036)	-0.196*** (0.033)	-0.129** (0.042)
学习策略	-0.072* (0.028)	-0.209*** (0.025)	-0.346*** (0.034)
监测	-0.067** (0.023)	-0.149*** 0.020)	-0.232*** (0.026)
编程	0.637*** (0.030)	0.623*** (0.022)	0.609*** (0.024)

表5：O*NET基本技能类别中每项技能的职业水平、人类注释的GPT暴露对技能重要性的回归，加上编程技能。这些技能的描述可以在附录B中找到。

工作需要的区域	**准备工作**	**教育需要**	**职业实例**	**中位数收入**	**雇员总数(000s)**	**H □**	**M □**	**H □**	**M □**	**H □**	**M □**
1	没有或很少(0-3个月)	高中文凭或普通教育证书(Otional)	食品准备工人。洗碗机、地板打磨机	$30,230	13,100	0.03	0.04	0.06	0.06	0.09	0.08
2	部分（3-12个月）	高中毕业证书	服务员、客户服务代表、出纳员	$38,215	73,962	0.07	0.12	0.16	0.20	0.24	0.27
3	中等（1-2年）	职业学校，在职培训，或副学士学位	电工、理发师、医疗助理	$54,815	37,881	0.11	0.14	0.26	0.32	0.41	0.51
4	相当长的时间（2-4年）	学士学位	数据库管理员、图形设计师、成本估算师	$77,345	56,833	0.23	0.18	0.47	0.51	0.71	0.85
5	广泛的（4年以上）	硕士或以上学位	药剂师、律师、天文学家	$81,980	21,221	0.23	0.13	0.43	0.45	0.63	0.76

表6：按就业区划分的GPT平均风险。对于每个工作区，我们还列出了每个构成职业的年收入中位

数（美元），以及该工作区所有职业的工人总数，以千为单位。

需要在职培训	收入中位数	雇员总数（千人）	H □	M □	H □	M □	H □	M □
无	$77,440	90,776	0.20	0.16	0.42	0.46	0.63	0.76
学徒制	$55,995	3,066	0.01	0.02	0.04	0.06	0.07	0.10
实习/居留	$77,110	3,063	0.16	0.06	0.36	0.38	0.55	0.71
短期在职培训	$33,370	66,234	0.11	0.15	0.21	0.25	0.32	0.34
适度的在职培训	$46,880	31,285	0.09	0.12	0.21	0.25	0.32	0.38
长期的在职培训	$48,925	5,070	0.08	0.10	0.18	0.22	0.28	0.33

表7：职业的平均接触分数，按达到工作能力所需的在职培训水平分组。除了暴露分数，我们还显示了每个职业的年收入中位数，以及每组工人的总数，以千计。

5 措施的验证

5.1 与早期工作的比较

本文旨在建立在以前的一些实证研究的基础上，研究人工智能和/或自动化进步的职业风险。以前的研究使用了各种方法，包括：1:

- 使用O*NET这样的职业分类法来描述哪些职业有常规与非常规、手工与认知任务的内容（Autor等人，2003；Acemoglu和Autor，2011a）。
- 将任务的文本描述与专利中的技术进步描述进行映射。(Kogan等人, 2021; Webb, 2020)
- 将人工智能系统的能力与职业能力联系起来，并将暴露估计值汇总到需要这些能力的职业中。(Felten等人, 2018, 2023)
- 通过一组从认知科学文献中提取的14种认知能力，将人工智能任务基准评估（ImageNet、Robocup等）的结果映射到59个工人任务。(Tolan等人，2021年)
- 专家对一组专家有高度信心的O*NET职业的自动化潜力进行标注，结合概率分类器来估计O*NET职业的其余部分的自动化潜力。(Frey和Osborne, 2017)
- 制定一个评估工人在经济中完成的活动 "是否适合机器学习"（SML）的评分标准（Brynjolfsson和Mitchell，2017；Brynjolfsson等人，2018，2023）。

我们在表8中提供了一组关于许多这些先前努力的汇总统计。

本文的方法主要建立在SML方法的基础上，开发了一个评分标准，以评估OLM能力与O*NET数据库中报告的工人任务之间的重叠。表9列出了我们新的LLM暴露测量值对来自（Felten等人，2018）（表中 "AI职业暴露得分"）、（Frey和Osborne，2017）（Frey和Osborne自动化）、（Webb，2020）中所有三种技术的得分、来自（Acemoglu和Autor，2011a）和（Brynjolfsson等人，2018，2023）（SML）的归一化常规手工和认知得分的OLS回归结果。我们还使用最新的BLS职业就业调查的年化职业工资作为控制。在本文中，有四个独立的输出变量代表新的分数，这些分数是由早期的努力预测的。

GPT-4 暴露评级 1 相当于我们用 GPT-4 评估的整体暴露评分标准，其中完全暴露的可能性被编码为 1，没有暴露的可能性被编码为 0，部分暴露（在我们的标签方案中为 E2）被编码为 0.5。GPT-4暴露等级2的评分与总体暴露相似，但提示略有不同。这两个提示的结果非常相似。GPT-4自动化评级应用了我们的 "T "评分标准，将没有来自LLM的自动化暴露编码为0，完全自动化暴露为1，2、3、4级分别为0.25、0.5和0.75。最后，人类暴露等级代表了与GPT-4暴露等级1相同的评分标准，但由人类来打分，这一点在本文的前一部分已经讨论过。这些结果与上面介绍的□组统计数据相对应。

每种测量方式的结果都是一致的。我们发现一般来说，正的和统计学上的
我们的LLM暴露测量值与以前针对软件和人工智能的测量值之间存在着明显的相关性。令人鼓舞

的是，按职业划分的SML暴露分数与我们在本文中开发的暴露分数显示出明显的正相关，表明这两项研究的方法相似，具有一定程度的凝聚力。基于Webb软件和人工智能专利的衡量标准，SML，以及归一化（去重）的

	闵行区	第25届Perc.	中位数	第75届Perc	最大	平均值	标准值。偏差。	计数
GPT-4暴露等级1	0.00	0.13	0.34	0.50	1.00	0.33	0.22	750
GPT-4暴露等级2	0.00	0.09	0.24	0.40	0.98	0.26	0.20	750
人体接触等级	0.00	0.09	0.29	0.47	0.84	0.29	0.21	750
软件(Webb)	1.00	25.00	50.00	75.00	100.00	50.69	30.05	750
机器人（Webb）	1.00	22.00	52.00	69.00	100.00	48.61	28.61	750
AI (Webb)	1.00	28.00	55.00	82.00	100.00	54.53	29.65	750
对机器学习的适用性	2.60	2.84	2.95	3.12	3.55	2.99	0.18	750
正常化的常规认知	-3.05	-0.46	0.10	0.63	3.42	0.07	0.86	750
正常化的常规手册	-1.81	-0.81	-0.11	0.73	2.96	0.05	1.01	750
AI职业暴露得分	1.42	3.09	3.56	4.04	6.54	3.56	0.70	750
弗雷和奥斯本自动化	0.00	0.07	0.59	0.88	0.99	0.50	0.38	681
平均数。工资	10.13	10.67	11.00	11.34	12.65	11.02	0.45	749

表8：先前测量人工智能和自动化职业暴露的一系列努力的汇总统计。我们还包括了本工作中新提出的测量的汇总统计。我们包括来自（Webb，2020）的所有措施，来自（Acemoglu和Autor，2011a）的归一化常规认知和手动得分（由于职业组的不完全匹配，平均值可能略微偏离0），来自（Brynjolfsson和Mitchell，2017；Brynjolfsson等，2018，2023）的机器学习适宜性，来自（Felten等，2018）的AI职业暴露，和来自（Frey和Osborne，2017）的自动化暴露。我们包括尽可能多的职业，但由于O*NET分类标准随着这些措施的制定而改变，一些角色可能在最新版本的O*NET 6位数职业中丢失。

并除以标准差）常规认知分数都与我们的一些措施表现出正相关。

软件、SML和常规认知分数都与LLM的接触分数在1%的水平上表现出正向和统计学意义上的关联。来自（Webb，2020）的AI分数的系数也是正的，并且在5%的水平上有统计学意义，但是我们在第3列和第4列中对LLM的整体暴露的二次提示没有表现出统计学意义上的关系。在大多数情况下，人工智能职业暴露得分与我们的暴露措施没有关联。Webb的机器人暴露得分、常规手工任务内容和来自（Frey和Osborne，2017）的整体自动化指标都与我们的主要GPT-4和人类评估的整体暴露评级呈负相关，以其他测量为条件。这种负相关反映了物理任务对LLM的有限暴露。手工作业暂时没有接触到LLM，甚至没有接触到有额外系统集成的LLM。我们的自动化评分结果也与（Frey和Osborne，2017）的衡量标准不相关。

与（Felten等人，2018）和（Frey和Osborne，2017）的低相关性可能是由方法的不同所解释的。将人工智能能力与工人的能力联系起来，或直接根据职业的特点对暴露进行评分，而不是从DWA或任务层面的评分汇总到职业（如SML的论文和我们自己的论文），为职业的内容提供了一个稍微不同的视角。

在所有的回归中，$\square^2$，在60.7%（第3列）和72.8%（第5列）之间。这表明，我们的衡量标准，明确关注LLM的能力，有28%到40%的未解释方差。

与其他测量方法相比。特别是在与人工智能有关的暴露分数的情况下，我们预计其他测量的组合将与我们的分数有很强的相关性。然而，早期的努力对LLM技术的未来进展信息有限。我们预计，我们对未来机器学习技术的理解也同样没有被我们今天的评分标准完美地捕捉到。

	GPT-4暴露等级		1GPT-4暴露等级		2人类暴露等级	
	(1)	(2)	(3)	(4)	(5)	(6)
软件(Webb)	0.00113***	0.00123***	0.00111***	0.00119***	0.00096***	0.00101***
	(0.00031)	(0.00031)	(0.00031)	(0.00031)	(0.00031)	(0.00031)
机器人（Webb）	-0.00378***	-0.00405***	-0.00377***	-0.00399***	-0.00371***	-0.00383***
	(0.00032)	(0.00031)	(0.00034)	(0.00033)	(0.00029)	(0.00028)
AI (Webb)	0.00080***	0.00090***	0.00036	0.00045	0.00067**	0.00071**
	(0.00030)	(0.00029)	(0.00030)	(0.00030)	(0.00030)	(0.00030)
对机器学习的适用性	0.29522***	0.26888***	0.28468***	0.26245***	0.19514***	0.18373***
	(0.04503)	(0.04418)	(0.04404)	(0.04342)	(0.03990)	(0.03886)
正常化的常规认知	0.06601***	0.06868***	0.04743***	0.05015***	0.03568***	0.03659***
	(0.00886)	(0.00894)	(0.00872)	(0.00879)	(0.00671)	(0.00669)
正常化的常规手册	-0.11147***	-0.11371***	-0.09390***	-0.09561***	-0.11045***	-0.11152***
	(0.00785)	(0.00789)	(0.00817)	(0.00818)	(0.00741)	(0.00744)
AI职业暴露得分	0.00993	0.02465**	-0.01537	-0.00265	0.00630	0.01252
	(0.01107)	(0.01059)	(0.01160)	(0.01114)	(0.00918)	(0.00845)
弗雷和奥斯本自动化	-0.03024*	-0.03950**	-0.00364	-0.01217	-0.03890**	-0.04253**
	(0.01835)	(0.01841)	(0.02007)	(0.01972)	(0.01883)	(0.01858)
平均数。工资	0.05804***		0.04863***		0.02531	
	(0.01870)		(0.01860)		(0.01727)	
恒定	-1.12937***	-0.45743***	-0.96117***	-0.39935***	-0.47078*	-0.17706
	(0.26859)	(0.15327)	(0.26365)	(0.15017)	(0.24684)	(0.13256)
N	680.00000	681.00000	680.00000	681.00000	680.00000	681.00000
□2	0.68741	0.68212	0.60737	0.60198	0.71213	0.71126

表9：GPT-暴露核心对先前努力的回归。从我们对早期量化职业暴露于人工智能和自动化的努力中的暴露措施的回归系数。我们还包括来自2021年5月BLS-OES调查的年化工资。每项措施都保持原来的规模，但来自（Acemoglu和Autor，2011a）的常规认知和常规手工得分除外。这两个分数被标准化为平均值为零，方差为1。一般来说，我们发现与以前的努力有很强的正相关，尽管大量的残余方差仍然可以由我们的新措施来解释。第1列和第2列是基于我们对GPT-4评级的主要□暴露测量。第3栏和第4栏是基于一个类似的、稍有不同的、也是由GPT-4评分的曝光度标准，以保证稳健性。第5栏和第6栏反映了人类在与第1栏和第2栏相同的评分标准上的评分。

6 讨论

6.1 GPT作为一种通用技术

在本文的前面，我们讨论了GPT可以被归类为一种通用技术的可能性。这种分类要求GPTs满足三个核心标准：随着时间的推移而改进，在整个经济中的普遍性，以及催生互补性创新的能力（Lipsey等人，2005）。来自人工智能和机器学习文献的证据彻底表明，GPTs满足了第一个标准--它们的能力随着时间的推移不断提高，有能力完成或帮助完成一系列日益复杂的任务和用例（见2.1）。本文提出了支持后两个标准的证据，发现GPT本身可以对整个经济产生普遍的影响，而GPT所带来的补充性创新--特别是通过软件和数字工具--可以广泛地应用于经济活动。

图3说明了建立在LLM之上的补充性软件的潜在经济影响。在X轴的某一点上，取□和□之间的Y轴之差（在所有职业中的份额）（在一个职业中被暴露的任务份额），可以得到职业内的总暴露潜力，超过了LLM本身的直接暴露。使用GPT-4注释的所有任务中，□和□的平均值为0.42，使用人类注释的平均值为0.32（见图3），这表明由GPT驱动的软件对任务暴露的平均影响可能比LLM本身的平均暴露大两倍以上（基于人类注释和GPT-4注释的平均值为0.14）。我们的研究结果表明，这些模型与相当份额的工人和任务有关，但它们也表明，它们所产生的软件创新可以推动更广泛的影响。

一项技术的普遍性的一个组成部分是它被企业和用户采用的程度。本文没有系统地分析这些模式的采用情况，但是，有早期的定性证据表明，对LLM的采用和使用正变得越来越广泛。在LLMs基础上进行的相对简单的UI改进的力量在ChatGPT的推出中显而易见--其中底层模型的版本以前是通过API提供的，但在ChatGPT界面发布后，使用率急剧上升。(Chow, 2023; OpenAI, 2022）在这次发布之后，一些商业调查表明，在过去几个月中，公司和工人对LLM的采用有所增加。(Constantz, 2023; ResumeBuilder.com, 2023)

然而，广泛地采用这些模型，就必须识别现有的瓶颈。决定其效用的一个关键因素是人类对它们的信任程度，以及习惯。例如，在法律界，这些模型的效用取决于法律专业人员是否能够信任它们的输出，而不需要求助于核实原始文件或进行独立研究。技术的成本和灵活性、工人和公司的偏好以及激励措施在采用建立在法律硕士之上的工具方面也起着重要作用。这样一来，采用的动力可能来自于与LLMs相关的一些道德和安全风险方面的进展：偏见、编造事实和错位，仅举几例OpenAI（2023a）。

此外，由于数据可用性、监管质量、创新文化以及权力和利益分配等因素，不同经济部门对大型语言模型的采用会有所不同。因此，要全面了解工人和公司对大型语言模型的采用和，需要对这些错综复杂的问题进行更深入的探讨。

一种可能性是，对于大多数任务来说，节省时间和无缝应用将比提高质量更重要。另一种情况是，最初的重点将是增强，然后是自动化（Huang和Rust，2018）。这可能会形成的一种方式是，在完全自动化之前，工作首先变得更不稳定（作家成为自由职业者）的增强阶段可能会出现。

6.2 对美国公共政策的影响

自动化技术的引入，包括LLMs，以前与经济差距的扩大和劳动力的中断有关，这可能会引起下游的不利影响。2021年；Klinova和Korinek，2021年；Weidinger等人，2021年，2022年）我们对美国工人接触的结果强调了社会和政策对LLM及其产生的补充技术所带来的潜在经济破坏的准备的必要性。虽然推荐具体的政策处方以平稳过渡到一个越来越广泛采用LLM的经济体不在本文的范围内，但之前的工作，如（Autor等人，2022b）已经阐明了美国政策的几个重要方向，涉及教育、工人培训、安全网计划的改革等等。

6.3 局限性和未来工作

这项研究有几个局限性，需要进一步调查。首先，我们对美国的关注限制了我们的研究结果在其他国家的推广，在这些国家，由于工业组织、技术基础设施、监管框架、语言多样性和文化背景等因素，生成性模型的采用和影响可能有所不同。我们希望通过扩大研究范围和分享我们的方法来解决这一局限性，以便其他研究人员能够在此基础上进行研究。

随后的研究工作应考虑两项额外的研究：一项是探索不同部门和职业的GPT采用模式，另一项是仔细研究最先进的模型与工人活动相关的实际能力和局限性，超出我们的暴露评分范围。例如，尽管最近GPT-4在多模态能力方面取得了进展，但我们并没有考虑视觉能力在直接GPT曝光的□评级。(OpenAI, 2023b)未来的工作应该考虑这种能力进步的影响，因为它们正在展开。我们承认，理论和实际表现之间可能存在差异，特别是在复杂、开放和特定领域的任务中。

7 总结

总之，本研究对法律硕士，特别是GPTs，对美国经济中各种职业和行业的潜在影响进行了研究。通过应用一个新的标准来理解LLM能力及其对工作的潜在影响，我们观察到大多数职业都在一定程度上暴露于GPTs，高工资的职业通常会出现更多的高暴露任务。我们的分析表明，当考虑到目前的模型能力和预期的GPT驱动的软件时，大约19%的工作有至少50%的任务暴露在GPT中。

我们的研究旨在强调GPTs的通用潜力及其对美国工人的可能影响。以前的文献表明，迄今为止，GPTs的改进令人印象深刻（见2.1）。我们的研究结果证实了这样的假设，即这些技术可以对美国广泛的职业产生普遍的影响，而且由GPTs支持的额外进步，主要是通过软件和数字工具，可以对一系列经济活动产生重大影响。然而，虽然GPTs使人类劳动更有效率的技术能力似乎很明显，但必须认识到社会、经济、监管和其他因素可能影响实际的劳动生产率结果。随着能力的不断发展，GPTs对经济的影响可能会持续和增加，给政策制定者预测和规范其发展轨迹带来了挑战。

进一步的研究是必要的，以探索GPT进步的更广泛的影响，包括它们增加或取代人类劳动的潜力，它们对工作质量的影响，对不平等的影响，技能发展，以及许多其他结果。通过寻求了解能力和潜在影响

通过了解GPT对劳动力的影响，政策制定者和利益相关者可以做出更明智的决定，以驾驭人工智能的复杂局面及其在塑造未来工作中的作用。

7.1 GPT结论（GPT-4的版本）

生成性预训练转化器（GPTs）产生了深刻的转变，获得了潜在的技术增长，渗透到任务中，极大地影响了职业。这项研究探测了GPTs的潜在轨迹，提出了一个开创性的标准来衡量任务的GPT暴露，特别是在美国劳动力市场。

7.2 GPT结论（作者增订版）

生成性预训练转化器（GPTs）产生了深刻的转变，获得了潜在的技术逻辑增长，渗透到任务中，消化了专业管理。衡量可能的轨迹？生成开创性的分类法，将决策者聚集在一起，归纳出今天的情况。

鸣谢

感谢帮助我们注释任务暴露的注释者小组，包括Muhammad Ahmed Saeed, Bongane Zitha, Merve Özen Şenen, J.J., 和Peter Hoeschele。我们还感谢Lauryn Fuld、 Ashley Glat、 Michael Lampe和Julia Susser的出色研究协助。我们感谢Miles Brundage对本文的重要反馈。

我们感谢Todor Markov和Vik Goel为我们建立了针对GPT-4运行分类法的基础设施。我们感谢Lama Ahmad, Donald Bakong, Seth Benzell, Erik Brynjolfsson, Parfait Eloundou-Enyegue, Carl Frey, Sarah Giroux, Gillian Hadfield, Johannes Heidecke, Alan Hickey, Eric Horvitz, Shengli Hu, Ashyana Kachra,Christina Kim, Katya Klinova, Daniel Kokotajlo, Gretchen Krueger, Michael Lampe, Aalok Mehta, Larissa Schiavo, Daniel Selsam, Sarah Shoker, Prasanna Tambe, 和 Jeff Wu 在项目的各个阶段提供了反馈和编辑。

法律硕士援助声明

在这个项目中，GPT-4和ChatGPT被用于写作、编码和格式化的协助。

A 分类法

A.1 曝光

E 曝光评分表

考虑一下最强大的OpenAI大型语言模型（LLM）。这个模型可以完成许多任务，这些任务可以被表述为有文本输入和文本输出，其中输入的上下文可以在2000字内捕获。该模型也不能得出

最新的事实（那些<1年前的事实），除非它们在输入中被捕获。

假设你是一名工作人员，在你的角色中具有平均水平的专业知识，试图完成给定的任务。你可以使用LLM以及任务中提到的任何其他现有软件或计算机硬件工具。你还可以使用任何可通过笔记本电脑获得的常用技术工具（例如，麦克风、扬声器等）。你不能使用任何其他物理工具或材料。

请根据下面的评分标准给给定的任务贴上标签。

同等质量意味着审查工作的人将无法分辨出是人类自己完成的还是在法学硕士的协助下完成的。

如果你不确定如何判断一项任务所需的时间，可以考虑所描述的工具是否暴露了与该任务相关的大多数子任务。

E1 - 直接接触

如果仅通过ChatGPT或OpenAI游乐场这样的接口直接访问LLM，就可以将完成同等质量的任务所需的时间至少减少一半，则将任务标记为E1。这包括可以减少到的任务。- 根据复杂的指令编写和转换文本和代码， - 按照规范对现有文本或代码提供编辑， - 编写代码，可以帮助执行过去由手工完成的任务， - 在语言之间翻译文本， - 总结中等长度的文件。
- 对文件提供反馈，- 回答关于文件的问题，- 生成用户可能想问的关于文件的问题，- 为面试或评估撰写问题，- 撰写和回复电子邮件，包括涉及反驳信息或参与谈判的电子邮件（但只有在谈判是通过书面通信的情况下），- 维护书面数据的记录，- 根据一般知识准备培训材料，或- 通过任何书面或口头媒介告知任何人任何信息。

E2--由LLM驱动的应用程序的曝光率

标签任务E2如果单单有机会使用LLM，可能不会将完成任务的时间至少减少一半，但很容易想象在LLM的基础上可以开发更多的软件，将完成任务的时间减少一半。这种软件可能包括以下功能。- 总结超过2000字的文件，并回答有关这些文件的问题， - 从互联网上检索最新的事实，并将这些事实与LLM的功能相结合。
- 检索组织的现有知识、数据或文件，并检索信息， - 检索高度专业化的领域知识， - 根据数据或书面输入提出建议， - 分析书面信息，为决策提供信息， - 根据高度专业化的知识编写培训材料，- 就问题提供咨询，以及 - 维护复杂的数据库。

E3 - 曝光给定的图像功能

假设你既能接触到LLM，又能接触到一个可以查看、说明和创建图像的系统，以及任何由LLM驱动的系统（上面E2中的系统）。这个系统不能将视频作为输入，也不能将视频作为输出。该系统不能准确地从图像输入中检索非常详细的信息，如图像中的尺寸测量。如果在使用LLM和这些图像能力的情况下，完成任务所需的时间明显减少，则将任务标记为E3。- 从PDF文件中阅读文本， - 扫描图像，或 - 根据指示创建或编辑数字图像。

图像可以是现实的，但不应该是详细的。模型可以识别图像中的物体，但不能识别这些选项之间的关系。

E0 - 没有接触

如果上述任何一项都不能使一个有经验的工人高质量地完成任务的时间至少减少一半，则将任务标记为E0。一些例子。- 如果一项任务需要高度的人际互动（例如，亲身示范），那么它应该被归为E0。- 如果一项任务需要精确的测量，那么它应该被列为E0级。- 如果一项任务需要详细审查视觉效果，那么它应该被归类为E0。- 如果一项任务需要使用手或走路，那么它应该被归类为E0。- 建立在LLM之上的工具不能做出任何可能影响人类生活的决定（例如，雇用、分级等）。如果任务的任何部分涉及到收集输入以做出最终决定（而不是分析数据以告知决定或提出建议），那

么它应该被归类为E0。LLM可以提出建议。- 即使建立在LLM之上的工具可以完成一项任务，如果使用这些工具不能为有经验的工人节省大量的时间来完成这项任务，那么它应该被列为E0级。- LLM和建立在它之上的系统不能做任何法律上需要人类来完成的任务。- 如果现有的技术不是由LLM驱动的，而且是常用的，可以完成该任务，那么如果使用LLM或由LLM驱动的工具不会进一步减少完成任务的时间，你应该将该任务列为E0级。

完成任务。

当有疑问时，你应该默认为E0。##

注释的例子。

职业。检查员、测试员、分拣员、取样员和称量员 任务。调整、清洁或修理产品或加工设备以纠正检查中发现的缺陷。标签（E0/E1/E2/E3）。E0 解释。该模型无法获得任何形式的实物，超过一半的任务（调整、清洁和修理设备）描述需要手或其他体现。

职业。计算机和信息研究科学家 任务。运用理论知识和创新来创造或应用新的技术，如调整应用计算机的原则，使其用于新的用途。标签（E0/E1/E2/E3）。E1 解释。模型可以在训练中学习理论知识，作为其一般知识库的一部分，而适应的原则可以在输入到模型的文本中捕获。

活动。安排餐饮预订。标签（E0/E1/E2/E3）。E2 解释。自动化技术已经存在（例如Resy），不清楚LLM在使用该技术的基础上能提供什么（无差异）。也就是说，你可以建立一些东西，允许你要求LLM为你在Resy上进行预订。

-

B ONET基本技能的定义

基本技能

发达的能力有助于学习或更迅速地获得知识。

内容

在各种不同领域的工作和获得更具体的技能所需的背景结构。

- **阅读理解** - 理解工作相关文件中的书面句子和段落。
- **积极倾听** - 全神贯注地听别人说话，花时间去理解别人的观点，适当地提出问题，不在不恰当的时候打断别人。
- **写作**--根据受众的需要，以书面形式进行有效沟通。
- **说话** - 与他人交谈，有效地传达信息。
- **数学** - 使用数学来解决问题。
- **科学** - 使用科学规则和方法来解决问题。

过程

有助于更迅速地获得各种领域的知识和技能的程序

- **批判性思维** - 使用逻辑和推理来确定替代解决方案、结论或解决问题的方法的优势和劣势。

- **主动学习**--了解新信息对当前和未来问题解决和决策的影响。

- **学习策略** - 在学习或教授新事物时，选择和使用适合情况的培训/教学方法和程序。
- **监测** - 监测/评估自己、其他个人或组织的表现，以做出改进或采取纠正措施。

跨职能的技能

注：我们只从跨职能的技能列表中选择了编程，因为我们事先了解了模型的编码能力。

- **编程** - 为各种目的编写计算机程序。

C 教育

	收入中位数	雇员（000人）	H □	M □	H □	M □	H □	M □
没有正式的教育证书	$31,900	36,187	0.05	0.06	0.10	0.10	0.15	0.15
高中文凭或同等学历	$45,470	67,033	0.09	0.13	0.20	0.25	0.31	0.37
中学后非学位奖	$48,315	9,636	0.07	0.15	0.19	0.28	0.31	0.41
一些大学，没有学位	$40,970	2,898	0.23	0.34	0.39	0.53	0.55	0.72
副学士学位	$60,360	3,537	0.12	0.14	0.31	0.36	0.49	0.59
学士学位	$78,375	71,698	0.23	0.17	0.47	0.51	0.70	0.84
硕士学位	$79,605	3,216	0.26	0.14	0.46	0.44	0.66	0.74
博士或专业学位	$82,420	5,290	0.21	0.13	0.41	0.43	0.60	0.74

表10：职业的平均暴露分数，按进入该职业所需的典型教育分组。除了暴露分数，我们还显示了每个职业的年收入中位数，以及每组工人的总数，以千计。

D 工业和生产力方面的风险

哪些地区最容易受到自动化和扩容的影响（地图）？

图6和图7分别显示了根据人类评分者和我们的算法风险评分标准，3位数NAICS行业的总体就业加权相对风险。几乎所有的行业都存在影响潜力，但有很大的异质性。两种方法在相对暴露上基本一致：数据处理、信息处理和医院都有高暴露。

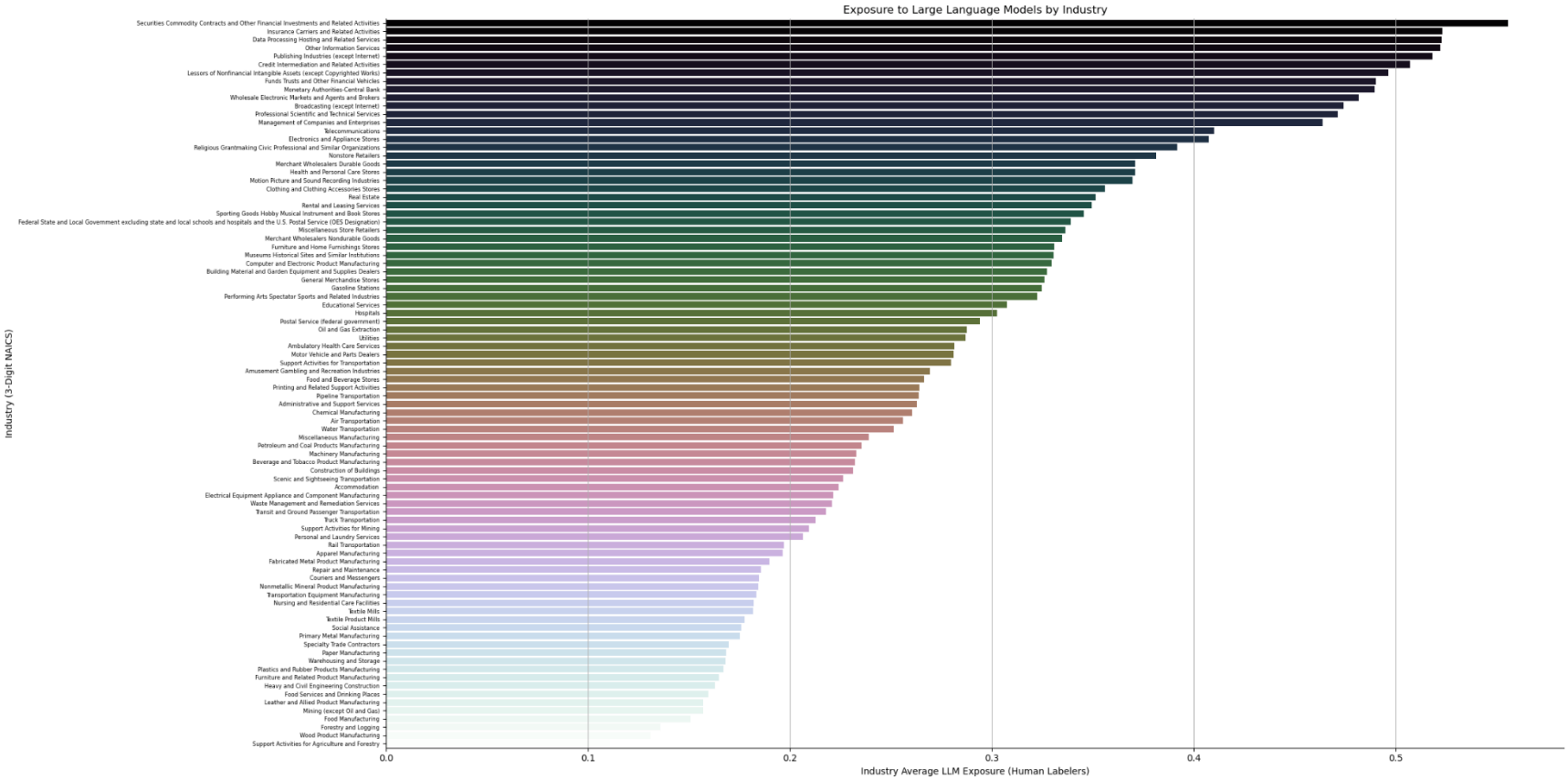
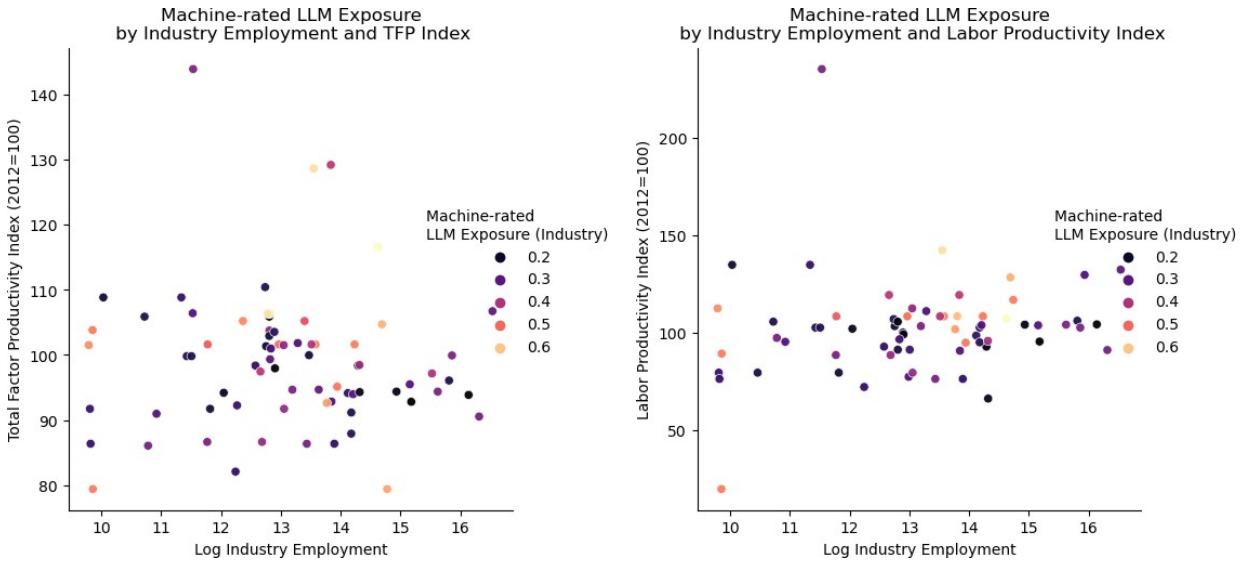

Exposure to Large Language Models by Industry
Industry (3-Digit NAICS)
Securities Commodity Contracts and Other Financial Investments and Related Activities
Insurance Carriers and Related Activities
Data Processing Hosting and Related Services
Other Information Services
Publishing Industries (except Internet)
Credit Intermediation and Related Activities
Lessors of Nonfinancial Intangible Assets (except Copyrighted Works)
Funds Trusts and Other Financial Vehicles
Monetary Authorities-Central Bank
Wholesale Electronic Markets and Agents and Brokers
Broadcasting (except Internet)
Professional Scientific and Technical Services
Management of Companies and Enterprises
Telecommunications
Electronics and Appliance Stores
Religious Grantmaking Civic Professional and Similar Organizations
Nonstore Retailers
Merchant Wholesalers Durable Goods
Health and Personal Care Stores
Motion Picture and Sound Recording Industries
Clothing and Clothing Accessories Stores
Real Estate
Rental and Leasing Services
Sporting Goods Hobby Musical Instrument and Book Stores
Federal State and Local Government excluding state and local schools and hospitals and the U.S. Postal Service (OES Designation)
Miscellaneous Store Retailers
Merchant Wholesalers Nondurable Goods
Furniture and Home Furnishings Stores
Museums Historical Sites and Similar Institutions
Computer and Electronic Product Manufacturing
Building Material and Garden Equipment and Supplies Dealers
General Merchandise Stores
Gasoline Stations
Performing Arts Spectator Sports and Related Industries
Educational Services
Hospitals
Postal Service (federal government)
Oil and Gas Extraction
Utilities
Ambulatory Health Care Services
Motor Vehicle and Parts Dealers
Support Activities for Transportation
Amusement Gambling and Recreation Industries
Food and Beverage Stores
Printing and Related Support Activities
Pipeline Transportation
Administrative and Support Services
Chemical Manufacturing
Air Transportation
Water Transportation
Miscellaneous Manufacturing
Petroleum and Coal Products Manufacturing
Machinery Manufacturing
Beverage and Tobacco Product Manufacturing
Construction of Buildings
Scenic and Sightseeing Transportation
Accommodation
Electrical Equipment Appliance and Component Manufacturing
Waste Management and Remediation Services
Transit and Ground Passenger Transportation
Truck Transportation
Support Activities for Mining
Personal and Laundry Services
Rail Transportation
Apparel Manufacturing
Fabricated Metal Product Manufacturing
Repair and Maintenance
Couriers and Messengers
Nonmetallic Mineral Product Manufacturing
Transportation Equipment Manufacturing
Nursing and Residential Care Facilities
Textile Mills
Textile Product Mills
Social Assistance
Primary Metal Manufacturing
Specialty Trade Contractors
Paper Manufacturing
Warehousing and Storage
Plastics and Rubber Products Manufacturing
Furniture and Related Product Manufacturing
Heavy and Civil Engineering Construction
Food Services and Drinking Places
Leather and Allied Product Manufacturing
Mining (except Oil and Gas)
Food Manufacturing
Forestry and Logging
Wood Product Manufacturing
Support Activities for Agriculture and Forestry
0.0
0.1
0.2
0.3
0.4
0.5
Industry Average LLM Exposure (Human Labelers)

图6

Exposure to Large Language Models by Industry
Industry (3-Digit NAICS)
Industry Average LLM Exposure (GPT-4)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Data Processing, Hosting, and Related Services
Other Information Services
Publishing Industries (except Internet)
Insurance Carriers and Related Activities
Credit Intermediation and Related Activities
Securities, Commodity Contracts, and Other Financial Investments and Related Activities
Lessors of Nonfinancial Intangible Assets (except Copyrighted Works)
Broadcasting (except Internet)
Professional, Scientific, and Technical Services
Funds, Trusts, and Other Financial Vehicles
Management of Companies and Enterprises
Wholesale Electronic Markets and Agents and Brokers
Telecommunications
Electronics and Appliance Stores
Nonstore Retailers
Religious, Grantmaking, Civic, Professional, and Similar Organizations
Computer and Electronic Product Manufacturing
Motion Picture and Sound Recording Industries
Merchant Wholesalers, Durable Goods
Publishing Industries
Federal, State, and Local Government excluding state and local schools and hospitals and the U.S. Postal Service (OES Designation)
Performing Arts, Spectator Sports, and Related Industries
Merchant Wholesalers, Nondurable Goods
Rental and Leasing Services
Printing and Related Support Activities
Clothing and Clothing Accessories Stores
Furniture and Home Furnishings Stores
Sporting Goods, Hobby, Musical Instrument, and Book Stores
Hospitals
Ambulatory Health Care Services
Oil and Gas Extraction
Miscellaneous Store Retailers
Museums, Historical Sites, and Similar Institutions
Support Activities for Transportation
Utilities
Pipeline Transportation
Educational Services
Building Material and Garden Equipment and Supplies Dealers
Motor Vehicle and Parts Dealers
Chemical Manufacturing
Transit and Ground Passenger Transportation
Couriers and Messengers
Truck Transportation
General Merchandise Stores
Water Transportation
Petroleum and Coal Products Manufacturing
Miscellaneous Manufacturing
Air Transportation
Administrative and Support Services
Electrical Equipment, Appliance, and Component Manufacturing
Waste Management and Remediation Services
Rail Transportation
Scenic and Sightseeing Transportation
Apparel Manufacturing
Support Activities for Mining
Fabricated Metal Product Manufacturing
Beverage and Tobacco Product Manufacturing
Couriers and Messengers
Gasoline Stations
Nonmetallic Mineral Product Manufacturing
Food Services and Drinking Places
Accommodation
Transportation Equipment Manufacturing
Textile Mills
Paper Manufacturing
Primary Metal Manufacturing
Textile Product Mills
Amusement, Gambling, and Recreation Industries
Personal and Laundry Services
Furniture and Related Product Manufacturing
Forestry and Logging
Plastics and Rubber Products Manufacturing
Specialty Trade Contractors
Heavy and Civil Engineering Construction
Leather and Allied Product Manufacturing
Repair and Maintenance
Mining (except Oil and Gas)
Warehousing and Storage
Nursing and Residential Care Facilities
Printing and Related Support Activities
Wood Product Manufacturing
Support Activities for Mining
Food Services and Drinking Places
Social Assistance

图7



最近的生产力增长（包括全要素和劳动力）似乎也与暴露不相关。图D和D显示，自2012年以来的生产力增长与模型评出的目前的LLMs风险之间没有什么关系。已经快速增长的生产性行业与暴露之间的高相关性可能意味着鲍莫尔成本病的加剧。换句话说，如果LLMs有可能在不同的行业中不同地提高生产力，那么一个令人担忧的问题是，生产力最高的行业会变得更加富有活力。由于对这些行业的生产需求缺乏弹性，最具生产力的部门在经济中的投入比例将会缩减。我们没有看到有什么迹象表明会出现这种情况。自2012年以来的生产力增长和接触LLM技术似乎没有关系。

E 没有任何暴露的任务的职业

没有标明暴露任务的职业
农业设备操作人员 运动员和体育竞技
者 汽车玻璃安装和维修人员
公共汽车和卡车技师及柴油发动机专家 水泥工和混凝土
粉刷工
厨师，快餐
切割机和修剪机，手动井架
操作人员，石油和天然气
餐厅和食堂服务员和酒保助手 洗碗工
挖泥机操作人员
电力线安装人员和维修人员
挖掘和装载机及拖曳机操作员，地面采矿业 地板铺设者，地毯、木材和
硬瓷砖除外
铸造厂的模具和型芯制造商
帮工-砖瓦工、砌块工、石匠、瓷砖和大理石镶嵌工 帮工-木匠
帮工-油漆工、裱糊工、抹灰工和泥瓦工 帮工-管道工、水
管工、管道钳工和蒸汽钳工 帮工-屋顶工
肉类、家禽和鱼类切割者和修剪者 摩托车机械
师
摊铺、铺面和夯实设备操作工 桩机操作工
浇注器和脚轮，金属
铁轨铺设和维护设备操作工 耐火材料维修工，砖匠
除外 采矿业屋顶螺栓工
抢劫者、石油和天然气屠宰
者和肉类包装者 石匠
锥子
轮胎维修工和更换工 井口抽
水机

表11：所有34种职业中，我们的测量方法都没有将任何任务标示为暴露。

参考文献

Abid, A., Farooqi, M., and Zou, J. (2021).大型语言模型中持续存在的反穆斯林偏见。在2021年

AAAI/ACM人工智能、伦理和社会会议的论文集中，AIES '21，第298-306页，美国纽约。计算机协会。

Acemoglu, D. (2002).技术变革、不平等和劳动力市场。经济文献杂志》，40。

Acemoglu, D. and Autor, D. (2011a).技能、任务和技术。对就业和收入的影响。载于《劳动经济学手册》，第4卷，第1043-1171页。Elsevier。

Acemoglu, D. and Autor, D. (2011b).技能、任务和技术。对就业和收入的影响。In Ashenfelter, O. and Card, D., editors, Handbook of Labor Economics, volume 4 of Handbook of Labor Economics, chapter 12, pages 1043-1171.Elsevier.

Acemoglu, D., Autor, D., Hazell, J., and Restrepo, P.（2020）。爱和工作。来自网上空缺职位的证据。技术报告，国家经济研究局。

Acemoglu, D. and Restrepo, P. (2018).人与机器之间的竞赛。技术对增长、要素份额和就业的影响。美国经济评论》，108（6）：1488-1542。

Acemoglu, D. and Restrepo, P. (2019).自动化和新任务。技术如何取代和恢复劳动力。经济展望》杂志，33（2）：3-30。

Acemoglu, D. and Restrepo, P. (2022a).人口统计学和自动化。经济研究评论》，89（1）：1-44。

Acemoglu, D. and Restrepo, P. (2022b).任务、自动化和美国工资不平等的上升。计量经济学报》，90（5）：1973-2016。

Agrawal, A. K., Gans, J. S., and Goldfarb, A. (2021).艾的采用和全系统的变化。技术报告，国家经济研究局。

Arntz, M., Gregory, T., and Zierahn, U. (2017).重新审视自动化的风险。经济学通讯》，159:157-160。

Autor, D., Chin, C., Salomons, A. M., and Seegmiller, B. (2022a).新的前沿阵地。新工作的起源和内容，1940-2018。技术报告，国家经济研究局。

Autor, D., Mindell, D. A., and Reynolds, E. B. (2022b).未来的工作。在智能机器的时代创造更好的工作。麻省理工学院出版社。

Autor, D. H., Katz, L. F., and Kearney, M. S. (2006).美国劳动力市场的两极分化。美国经济评论》，96（2）：189-194。

Autor, D. H., Levy, F., and Murnane, R. J. (2003).近期技术变革的技能含量。实证探索。经济学季刊》，118（4）：1279-1333。

Babina, T., Fedyk, A., He, A., and Hodson, J. (2021).人工智能、公司增长和产品创新。公司增长，和产品创新（2021年11月9日）。

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N. 。 Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T., Clark, J., McCandlish, S., Olah, C., Mann, B., and Kaplan, J. (2022).通过对人类反馈的强化学习来训练一个有用和无害的助手。 arXiv:2204.05862 [cs].

Baumol, W. J. (2012).成本病。为什么计算机越来越便宜，而医疗服务却不便宜。耶鲁大学出版社。

Benzell, S. G., Kotlikoff, L. J., LaGarda, G., and Ye, V. Y. (2021).模拟内生的全球自动化。工作文件29220，国家经济研究局。

Bessen, J. (2018).人工智能和就业。需求的作用。在《人工智能的经济学：一个议程》中，第291-307页。芝加哥大学出版社。

BLS (2022).按详细职业分类的就业。BLS

(2023a).人口特征（cps）。

BLS (2023b).职业前景手册a-z索引。

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021).论基础模型的机会和风险。arXiv预印本arXiv:2108.07258。

Bresnahan, T. (2019).人工智能技术和总量增长前景。

Bresnahan, T., Greenstein, S., Brownstone, D., and Flamm, K.（1996）。计算和计算机用途中的技术进步和共同发明。布鲁金斯经济活动论文。微观经济学，1996：1-83。

Bresnahan, T. F. (1999).计算机化和工资分散：一个分析性的重新解释。经济杂志》，109（456）：390-415。

Bresnahan, T. F., Brynjolfsson, E., and Hitt, L. M. (2002).信息技术、工作场所组织和对熟练劳动力的需求。公司层面的证据。经济学季刊》，117（1）：339-376。

Bresnahan, T. F. and Trajtenberg, M. (1995).通用技术的"增长引擎"？ Journal of econometrics, 65(1):83-108.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. （2020）。语言模型是很少的学习者。神经信息处理系统的进展，33：1877-1901。

Brynjolfsson, E., Frank, M. R., Mitchell, T., Rahwan, I., and Rock, D. (2023).量化机器学习对工作影响的分布。即将出版。

Brynjolfsson, E. and Mitchell, T. (2017).机器学习能做什么？ 劳动力影响。科学》，358（6370）：1530-1534。

Brynjolfsson, E., Mitchell, T., and Rock, D. (2018).机器可以学习什么，这对职业和经济意味着什么？AEA论文和会议记录，108：43-47。

Brynjolfsson, E., Rock, D., and Syverson, C. (2021).生产力J型曲线。无形资产如何补充通用技术。美

国经济杂志。宏观经济学，13（1）：333-72。

Chase, H. (2022).LangChain.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al. （2021）。评估在代码上训练的大型语言模型。arXiv预印本arXiv:2107.03374。

Cheng, Z., Lee, D., and Tambe, P. (2022).Innovae：用于理解专利和创新的生成性ai。可在SSRN查阅。

Chow, A. R. (2023).为什么ChatGPT是有史以来增长最快的网络平台|时间。

Cockburn, I. M., Henderson, R., and Stern, S. (2018).人工智能对创新的影响。一个探索性的分析。在《人工智能的经济学。一个议程，第115-146页。芝加哥大学出版社。

Constantz, J. (2023).根据一项新的调查，近三分之一的白领工人曾尝试过chatgpt或其他ai项目。

David, P. A. (1990).发电机和计算机：现代生产力悖论的历史视角。美国经济评论》，80（2）：355-361。

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019).Bert：用于语言理解的深度双向变换器的预训练。ArXiv，abs/1810.04805。

Dixon, J., Hong, B., and Wu, L. (2021).机器人革命。公司的管理和就业后果。管理科学》，67（9）：5586-5605。

Feigenbaum, J. J. and Gross, D. P. (2021).组织摩擦和自动化收益的增加。二十世纪AT&T的教训。技术报告，国家经济研究局。

Felten, E., Raj, M., and Seamans, R. (2023).arXiv预印本arXiv:2303.01157。

Felten, E. W., Raj, M., and Seamans, R. (2018).将人工智能的进展与职业能力联系起来的方法。AEA论文和会议记录，108:54-57。

Frey, C. B. (2019).技术陷阱。In The Technology Trap.普林斯顿大学出版社。Frey, C. B. and Osborne, M. A. (2017).就业的未来。工作对计算机化的影响有多大？技术预测和社会变革, 114(C):254-280.

Goldfarb, A., Taska, B., and Teodoridis, F. (2023).机器学习可能是一种通用技术吗？ 利用在线招聘信息的数据对新兴技术进行比较。研究政策，52（1）：104653。

Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., and Sedova, K.（2023）。生成语言模型和自动影响操作。新出现的威胁和潜在的缓解措施。

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., and Evans, O.（2018）。来自AI专家的证据，AI何时会超过人类的表现？人工智能研究杂志》，62：729-754。

Hernandez, D., Kaplan, J., Henighan, T., and McCandlish, S. (2021).转移的缩放规律。arXiv预印本arXiv:2102.01293。

Horton, J. J. (2023).作为模拟经济主体的大型语言模型。What can we learn from homo silicus? arXiv preprint arXiv:2301.07543.

Huang, M.-H. and Rust, R. T. (2018).服务中的人工智能。服务研究杂志，21（2）：155-172。

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020).神经语言模型的缩放规律。arXiv预印本arXiv:2001.08361。

Katz, L. F. and Murphy, K. M. (1992).相对工资的变化，1963-1987年：供应和需求因素。经济学季刊》，107（1）：35-78。

Khlaaf, H., Mishkin, P., Achiam, J., Krueger, G., and Brundage, M. (2022).代码合成大型语言模型的危险分析框架。

Klinova, K. and Korinek, A. (2021).艾与共享繁荣。AIES 2021--2021年AAAI/ACM人工智能、伦理和社会会议记录。

Kogan, L., Papanikolaou, D., Schmidt, L. D. W., and Seegmiller, B. (2021).技术、特定年份的人力资本和劳动力转移。将专利与职业联系起来的证据。工作文件29552，国家经济研究局。

Korinek, A. (2023).用于经济研究的语言模型和认知自动化。技术报告，国家经济研究局。

Korinek, A. and Stiglitz, J. E. (2018).人工智能及其对收入分配和失业的影响。在《人工智能的经济学。一个议程，第349-390页。芝加哥大学出版社。

Lipsey, R. G., Carlaw, K. I., and Bekar, C. T.（2005）。经济转型：通用技术和长期经济增长。Oup Oxford.

Meindl, B., Frank, M. R., and Mendonça, J. (2021).职业对第四次工业革命技术的暴露。 arXiv预印本 arXiv:2110.13317。

Mialon, G., Dessì, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R., Rozière, B., Schick, T., Dwivedi-Yu, J., Celikyilmaz, A., etc. (2023).Augmented language models: a survey. arXiv preprint arXiv:2302.07842.

Moll, B., Rachel, L., and Restrepo, P. (2021).不平衡的增长。自动化对收入和财富不平等的影响。SSRN电子期刊。

Mollick, E. R. and Mollick, L. (2022).由ai聊天机器人实现的新的学习模式。三种方法和作业。可在SSRN查阅。

Noy, S. and Zhang, W. (2023).生成式人工智能的生产力效应的实验证据。可在SSRN 4375283查阅。

O*NET（2023）。O*net 27.2数据库。

OpenAI（2022）。介绍chatgpt。

OpenAI（2023a）。Gpt-4系统卡。技术报告，OpenAI。OpenAI

(2023b).Gpt-4技术报告。技术报告，OpenAI。

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). 训练语言模型以遵循人类反馈的指令。arXiv预印本 arXiv:2203.02155。

Peng, S., Kalliamvakou, E., Cihon, P., and Demirer, M. (2023).ai对开发者生产力的影响。Evidence from github copilot. arXiv preprint arXiv:2302.06590.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. （2019）。语言模型是无监督的多任务学习者。OpenAI博客，1（8）：9。

ResumeBuilder.com（2023）。1/4的公司已经用chatgpt取代了工人。

Rock, D. (2019).工程价值。技术人才和人工智能投资的回报。可在SSRN 3427412查阅。

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., and Scialom, T.（2023）。工具形成者。语言模型可以教自己使用工具。arXiv预印本arXiv:2302.04761。

Schramowski, P., Turan, C., Andersen, N., Rothkopf, C. A., and Kersting, K. (2022).大型预训练的语言模型包含类似人类的偏见，即什么是正确的，什么是错误的行为。Nature Machine Intelligence, 4(3):258-268.

Shahaf, D. and Horvitz, E. (2010).人类和机器计算的通用任务市场。AAAI人工智能会议论文集》。

Singla, A. K., Horvitz, E., Kohli, P., and Krause, A.（2015）。学习雇用团队。在AAAI人类计算与众包会议上。

Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Kim, J.W., Kreps, S., McCain, M., Newhouse, A., Blazakis, J., McGuffie, K., and Wang, J. (2019).发布策略和语言模型的社会影响。

Sorensen, T., Robinson, J., Rytting, C., Shaw, A., Rogers, K., Delorey, A., Khalil, M., Fulda, N., and Wingate, D.(2022).一个信息理论的方法来提示工程没有地面真实标签。载于《计算语言学协会第60届年会论文集》（第一卷：长篇论文）。计算语言学协会。

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., et al.（2022）。Lamda。用于对话应用的语言模型。arXiv预印本arXiv:2201.08239。

Tolan, S., Pesole, A., Martínez-Plumed, F., Fernández-Macías, E., Hernández-Orallo, J., and Gómez, E.（2021）。测量AI的职业影响：任务、认知能力和AI基准。人工智能研究杂志》，71：191-236。

Van Reenen, J. (2011).工资不平等、技术和贸易：21世纪的证据。劳动经济学，18（6）：730-741。

Webb, M. (2020).人工智能对劳动力市场的影响。工作文件，斯坦福大学。Weidinger, L. et al. (2021).

语言模型伤害的伦理和社会风险。 arXiv:2112.04359 [cs]。

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S. 。Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., Haas, J., Legassick, S., Irving, G., and Gabriel, I. (2022).语言模型所带来的风险分类学。In 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22, page 214-229, New York, NY, USA.计算机械协会。

Zolas, N., Kroff, Z., Brynjolfsson, E., McElheran, K., Beede, D. N., Buffington, C., Goldschlag, N., Foster, L., and Dinlersoz, E. （2021）。美国公司对先进技术的采用和使用。来自年度商业调查的证据。技术报告，国家经济研究局。