GPT 就是 GPT：对劳动力市场影响的早期观察
大型语言模型的潜力

蒂娜·埃伦杜1个, 萨曼宁1,2, 帕梅拉·米什金∗1个, 和丹尼尔洛克3个

1个开放人工智能
2个开放研究
3个宾夕法尼亚大学

2023 年 3 月 27 日

抽象的

我们调查了大型语言模型 (LLM)（例如生成式预训练转换器 (GPT)）对美国劳动力市场的潜在影响，重点关注与 LLM 本身相比，LLM 支持的软件所带来的能力提升。我们使用新的规则，根据职业与 LLM 能力的一致性来评估职业，同时整合人类专业知识和 GPT-4 分类。我们的调查结果显示，大约 80% 的美国劳动力至少有 10% 的工作任务会受到 LLM 的引入影响，而大约 19% 的员工可能会看到至少 50% 的工作任务受到影响。我们不对此类 LLM 的开发或采用时间表做出预测。预计的影响涵盖所有工资水平，高收入工作可能面临更多的 LLM 能力和 LLM 支持的软件。值得注意的是，这些影响并不局限于近期生产率增长较高的行业。我们的分析表明，通过获得 LLM，美国大约 15% 的工人任务可以在相同质量水平下更快地完成。当合并构建在 LLM 之上的软件和工具时，这一份额增加到所有任务的 47% 到 56%。这一发现意味着 LLM 支持的软件将对扩展基础模型的经济影响产生重大影响。我们得出结论，像 GPT 这样的 LLM 表现出通用技术的特征，表明它们可能具有相当大的经济、社会和政策影响。在美国，大约 15% 的工人任务可以在相同的质量水平下更快地完成。当合并构建在 LLM 之上的软件和工具时，这一份额增加到所有任务的 47% 到 56%。这一发现意味着 LLM 支持的软件将对扩展基础模型的经济影响产生重大影响。我们得出结论，像 GPT 这样的 LLM 表现出通用技术的特征，表明它们可能具有相当大的经济、社会和政策影响。在美国，大约 15% 的工人任务可以在相同的质量水平下更快地完成。当合并构建在 LLM 之上的软件和工具时，这一份额增加到所有任务的 47% 到 56%。这一发现意味着 LLM 支持的软件将对扩展基础模型的经济影响产生重大影响。我们得出结论，像 GPT 这样的 LLM 表现出通用技术的特征，表明它们可能具有相当大的经济、社会和政策影响。这一发现意味着 LLM 支持的软件将对扩展基础模型的经济影响产生重大影响。我们得出结论，像 GPT 这样的 LLM 表现出通用技术的特征，表明它们可能具有相当大的经济、社会和政策影响。这一发现意味着 LLM 支持的软件将对扩展基础模型的经济影响产生重大影响。我们得出结论，像 GPT 这样的 LLM 表现出通用技术的特征，表明它们可能具有相当大的经济、社会和政策影响。

1 简介

如图 1 所示，最近几年、几个月和几周在生成 AI 和大型语言模型 (LLM) 领域取得了显着进展。虽然公众经常将 LLM 与生成式预训练变压器 (GPT) 的各种迭代联系起来，但 LLM 可以使用一系列架构进行训练，并不局限于基于变压器的模型（Devlin 等人，2019 年）。LLM 可以处理和生成各种形式的顺序数据，包括汇编语言、蛋白质序列和国际象棋游戏，超越了自然语言应用本身。在本文中，我们可以互换使用 LLM 和 GPT，并在我们的规则中指定这些应该被视为类似于通过 ChatGPT 或 OpenAI Playground 提供的 GPT 系列模型（在标记时 GPT 中包含模型- 3. 5 家族但不在 GPT-4 家族中）。我们检查具有文本和代码生成能力的 LLM，使用术语“生成 AI”来额外包括图像或音频等模式，并使用“LLM 支持的软件”来涵盖构建在 LLM 之上或将 LLM 与其他生成式 AI 模型。

∗通讯作者（pamela@openai.com ）。作者贡献均等，并按字母顺序排列。

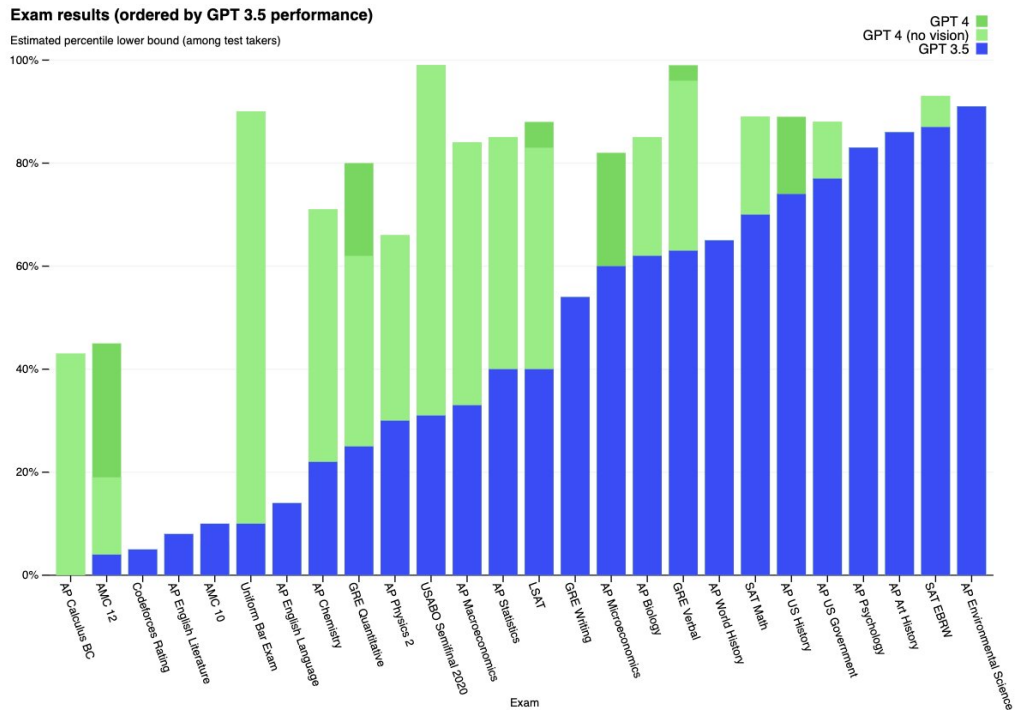


图 1：要了解模型功能的进步速度——考虑 GPT-3.5 和 GPT-4 之间考试成绩的跃升（OpenAI，2023b）。

不过，我们研究的动机并非仅受这些模型的进步影响，而更多的是受我们在围绕它们开发的互补技术中看到的广度、规模和能力的推动。互补技术的作用仍有待观察，但最大限度地发挥 LLM 的影响似乎取决于将它们与更大的系统集成（Bresnahan，2019 年；Agrawal 等人，2021 年）。虽然我们讨论的重点主要是 LLM 的生成能力，但重要的是要注意这些模型也可以用于文本生成以外的各种任务。例如，来自 LLM 的嵌入可用于自定义搜索应用程序，LLM 可以执行摘要和分类等任务，其中上下文可能主要包含在提示中。

为了补充对技术对工作影响的预测，并提供一个框架来理解语言模型及其相关技术不断发展的前景，我们提出了一个新的标准来评估 LLM 能力及其对工作的潜在影响。本准则 (A.1) 衡量任务对 LLM 的总体暴露程度，遵循先前量化机器学习暴露程度的工作精神（Brynjolfsson 等人，2018 年；Felten 等人，2018 年；Webb，2020 年）。我们将风险敞口定义为潜在经济影响的代表，而不区分劳动力增强或劳动力替代效应。我们使用人工注释器和 GPT-4 本身作为分类器，将此规则应用于主要来自 O*NET 数据库的美国经济中的职业数据。12

为了构建我们的主要曝光数据集，我们收集了人工注释和 GPT-4 分类，使用提示调谐以与作者的标签样本一致。当聚合到任务级别时，我们在 GPT-4 响应以及人机评估之间观察到类似的一致性级别。

1个这与最近使用 LLM 模拟人类行为的社会科学研究不同（Horton，2023；Sorensen 等人，2022）

2个虽然我们的曝光量规不一定将语言模型的概念与任何特定模型联系起来，但我们对 GPT-4 的观察能力以及我们在与 OpenAI 的启动合作伙伴（OpenAI，2023b）的开发中看到的功能套件产生了强烈的动机。

这种暴露量度反映了对提高人类劳动效率的技术能力的估计；然而，社会、经济、监管和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，在考虑当前模型功能和基于它们构建的预期工具时，大约 19% 的工作暴露了至少 50% 的任务。人类评估表明，在考虑现有的语言和代码能力而无需额外的软件或模式时，只有 3% 的美国工人将超过一半的任务暴露给 LLM。考虑到其他生成模型和补充技术，我们的人类估计表明，多达 49% 的工人可能有一半或更多的任务暴露于 LLM。

我们的研究结果一致表明，在人类和 GPT-4 注释中，大多数职业都表现出一定程度的 LLM 暴露，不同类型工作的暴露水平不同。工资较高的职业通常具有较高的接触率，这一结果与对机器学习总体接触率的类似评估相反（Brynjolfsson 等人，2023 年）。当使用 O*NET 的技能量表对技能集的曝光度测量进行回归时，我们发现严重依赖科学和批判性思维技能的角色与曝光度呈负相关，而编程和写作技能与 LLM 曝光度呈正相关。跟随 Autor 等人。(2022a)，我们检查了“工作区”的进入壁垒，发现 LLM 的职业暴露随着工作准备的难度而微弱增加。换句话说，

我们进一步将我们的测量结果与之前记录经济中自动化风险分布的努力进行比较，并发现大致一致的结果。我们检查的大多数其他技术暴露指标在统计 上与我们首选的暴露指标显着相关，而手动例行性和机器人技术暴露指标显示出负相关。这些早期努力解释的差异（Acemoglu 和 Autor，2011a；Frey 和 Osborne，2017；Brynjolfsson 等，2018；Felten 等，2018；Webb，2020；Brynjolfsson 等，2023），以及工资控制，范围从 60% 到 72%，这表明我们的 AI 暴露测量中有 28% 到 40% 的变化仍然没有被以前的技术暴露测量所解释。

我们按行业分析风险敞口，发现信息处理行业（4 位 NAICS）表现出高风险敞口，而制造业、农业和采矿业则表现出较低的风险敞口。过去十年的生产力增长与整体 LLM 暴露之间的联系似乎很弱，这表明一个潜在的乐观案例表明，LLM 未来生产力的提高可能不会加剧可能的成本疾病效应（Baumol，2012 年；Aghion 等人，2018 年）。3个

我们的分析表明，像 GPT-4 这样的 LLM 的影响可能是无处不在的。虽然法学硕士的能力随着时间的推移不断提高，但即使我们今天停止新能力的开发，它们不断增长的经济影响预计也会持续并增加。我们还发现，当我们考虑互补技术的发展时，LLM 的潜在影响会显着扩大。总的来说，这些特征意味着生成式预训练变压器 (GPT) 是通用技术 (GPT)。4 (Bresnahan 和 Trajtenberg，1995 年；Lipsey 等人，2005 年）。

(Goldfarb et al., 2023) 认为机器学习作为一个广泛的类别可能是一种通用技术。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状态的标准。本文的主要贡献是提供一组 LLM 影响潜力的测量，并演示应用 LLM 高效、大规模地开发此类测量的用例。此外，我们展示了 LLM 的通用潜力。如果“GPTs 就是 GPTs”，LLM 发展和应用的最终轨迹对于政策制定者来说可能难以预测和监管。与其他通用技术一样，这些算法中的大部分

3个Baumol 的成本病是一种理论，它解释了为什么医疗保健和教育等劳动密集型服务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熟练工人的工资增加了，但这些服务行业的生产率或效率没有相应的提高。因此，与经济中的其他商品和服务相比，这些行业的劳动力成本变得相对昂贵。

4个对于本文的其余部分，我们会在声明“GPT 是 GPT”之外使用时详细说明通用技术。

潜力将出现在广泛的具有经济价值的用例中，包括创造新的工作类型（Acemoglu 和 Restrepo，2018 年；Autor 等人，2022a）。我们的研究用于衡量现在技术上可行的东西，但必然会错过 LLM 随着时间的推移不断变化的影响潜力。

本文结构如下：第 2 节回顾相关的先前工作，第 3 节讨论方法和数据收集，第 4 节介绍汇总统计数据和结果，第 5 节将我们的测量与早期工作联系起来，第 6 节讨论结果，第 7 节提供结论评论。

2。文献综述

2.1 大型语言模型的进步

近年来，生成式 AI 模型因其处理各种复杂的基于语言的任务的能力而受到人工智能 (AI) 研究界和公众的极大关注。这些模型能力的进步受到多种因素的推动，包括模型参数数量的增加、训练数据量的增加和训练配置的增强（Brown 等人，2020 年；Radford 等人，2019 年；Hernandez 等人，2021 年）；卡普兰等人，2020 年）。广泛的、最先进的 LLM，例如 LaMDA（Thoppilan 等人，2022 年）和 GPT-4（OpenAI，2023b），在翻译、分类、创意写作和代码生成等各种应用程序中表现出色——能力以前需要专家工程师使用特定领域的数据开发专门的、任务特定的模型。

同时，研究人员使用微调和强化学习以及人类反馈等方法提高了这些模型的可控性、可靠性和实用性（Ouyang 等人，2022 年；Bai 等人，2022 年）。这些进步增强了模型辨别用户意图的能力，使它们更加用户友好和实用。此外，最近的研究揭示了 LLM 在编程和控制其他数字工具方面的潜力，例如 API、搜索引擎，甚至其他生成式 AI 系统（Schick 等人，2023 年；Mialon 等人，2023 年；Chase，2022 年）。这使得各个组件能够无缝集成，以实现更好的实用性、性能和泛化。在他们的极限下，这些趋势暗示了一个世界，在这个世界中，法学硕士可能能够执行通常在计算机上执行的任何任务。

生成式 AI 模型大多被部署为模块化专家，执行特定任务，例如从字幕生成图像或从语音转录文本。然而，我们认为将 LLM 视为用于创建其他工具的多功能构建块是必不可少的。开发这些工具并将它们集成到系统中将需要时间，并且可能需要对各个行业的现有流程进行重大重新配置。尽管如此，我们已经见证了新兴的采用趋势。尽管存在局限性，LLM 越来越多地被整合到写作辅助、编码和法律研究等领域的专业应用程序中。然后，这些专门的应用程序允许企业和个人将 LLM 应用到他们的工作流程中。

我们强调这些互补技术的重要性，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 LLM 可能继续对各种任务不可靠（Abid 等人，2021 年；Schramowski 等人，2022 年；Goldstein 等人）., 2023；OpenAI，2023a)。然而，专门的工作流程——包括工具、软件或人在回路系统——可以通过结合特定领域的专业知识来帮助解决这些缺点。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，利用嵌入和摘要来应对 GPT-4 可能提供有关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，它使用 LLM 生成代码片段和自动完成代码，然后用户可以根据自己的专业知识接受或拒绝。换句话说，虽然 GPT-4 本身确实“不知道现在几点”，但给它一块手表很容易。

此外，当 LLM 超过特定的性能阈值时，可能会出现一个正反馈循环，使他们能够协助构建能够增强其在各种领域的实用性和可用性的工具。

上下文。这可以降低创建此类工具所需的成本和工程专业知识，从而有可能进一步加速 LLM 的采用和集成（Chen 等人，2021 年；Peng 等人，2023 年）。LLM 还可以成为机器学习模型开发中的宝贵资产——充当研究人员、数据标签服务或合成数据生成器的编码助手。此类模型有可能有助于任务级别的经济决策，例如，通过改进人机之间任务和子任务分配的方法（Singla 等人，2015 年；Shahaf 和 Horvitz，2010 年）。随着 LLM 随着时间的推移而进步并更好地符合用户偏好，我们可以预期性能会不断提高。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 年；Weidinger 等人，2022 年；索莱曼等人，2019）

2.2 自动化技术的经济影响

大量且不断增长的文献探讨了人工智能和自动化技术对劳动力市场的影响。技能偏向技术变革的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的研究（Katz 和 Murphy，1992 年）. 许多研究都建立在这一概念之上，探索技术变革和自动化在基于任务的框架内对工人的影响（Autor 等人，2003 年；Acemoglu 和 Autor，2011b；Acemoglu 和 Restrepo，2018 年）。这方面的研究表明，从事日常和重复性工作的工人因技术驱动而流离失所的风险更高，一种被称为常规偏向技术变革的现象。最近的研究区分了技术的任务置换和任务恢复效应（新技术增加了对更广泛的劳动密集型任务的需求）（Acemoglu 和 Restrepo，2018 年，2019 年）。多项研究表明，自动化技术导致了美国的工资不平等，原因是专门从事日常工作的工人的工资相对下降（Autor 等人，2006 年；Van Reenen，2011 年；Acemoglu 和 Restrepo，2022b）。

先前的研究采用了多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。这些方通过认知能力对工人任务进行基准评估（Tolan 等人，2021 年），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这种潜力（Frey 和 Osborne，2017 年），建模任务级自动化并将结果聚合到职业级见解（Arntz 等人，2017 年），收集专家预测（Grace 等人，2018 年），与本文最相关的是，设计了一个新的规则来评估工人活动是否适合机器学习（Brynjolfsson 等人，2018 年，2023 年）。其中一些方法发现，在任务层面接触人工智能技术在职业中往往是多样化的。将每项工作视为一揽子任务，很难找到人工智能工具几乎可以完成所有工作的职业。(Autor et al., 2022a) 还发现自动化和增强暴露往往呈正相关。还有越来越多的研究检查 LLM 的特定经济影响和机会（Bommasani 等人，2021 年；Felten 等人，2023 年；Korinek，2023 年；Mollick 和 Mollick，2022 年；Noy 和 Zhang，2023 年；Peng 等人., 2023). 除了这项工作，

通用技术（例如印刷、蒸汽机）的特点是广泛传播、持续改进和互补创新的产生（Bresnahan 和 Trajtenberg，1995 年；Lipsey 等人，2005 年）。它们的深远影响将持续数十年，难以预测，尤其是在劳动力需求方面（Bessen，2018 年；Korinek 和 Stiglitz，2018 年；Acemoglu 等人，2020 年；Benzell 等人，2021 年）。实现通用技术的全部潜力需要广泛的共同发明（Bresnahan 和 Trajtenberg，1995 年；Bresnahan 等人，1996 年，2002 年；Lipsey 等人，2005 年；Dixon 等人，

2021 年），这是一个涉及发现新业务流程的昂贵且耗时的过程（David，1990 年；Bresnahan，1999 年；Frey，2019 年；Brynjolfsson 等人，2021 年；Feigenbaum 和 Gross，2021 年）。因此，许多关于机器学习技术的研究都侧重于系统级采用，认为组织系统可能需要重新设计才能有效利用新的机器学习进步（Bresnahan，2019 年；Agrawal 等人，2021 年；Goldfarb 等人，2023 年）. 设计得当的系统可以产生可观的商业价值并提高公司绩效（Rock，2019 年；Babina 等人，2021 年；Zolas 等人，2021 年），人工智能工具可促进发现过程（Cockburn 等人，2018 年；Cheng 等人., 2022). 通过使用任务级信息来评估 LLM 是否满足通用技术的标准，

我们试图以多种方式建立在这些不同的文献流之上。呼应（Felten 等人，2023 年），我们将分析重点放在 LLM 的影响上，而不是更广泛地解决机器学习或自动化技术。此外，我们提出了一种使用 LLM（特别是 GPT-4）的新方法来评估任务的曝光和自动化潜力，从而加强人工评分工作。随后，我们将我们的调查结果汇总到职业和行业，捕捉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 美国职业活动和任务的数据

我们使用 O*NET27.2 数据库 (O*NET, 2023)，它包含 1,016 个职业的信息，包括他们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，例如“研究脚本以确定项目要求”。另一方面，任务是特定于职业的工作单元，可能与零个、一个或多个 DWA 相关联。我们在表 1 中提供了任务和 DWA 的示例。我们使用的两个数据集包括：

- 19,265 个任务，包括“任务描述”和相应的职业，以及
- 2,087 个 DWA，其中大多数 DWA 连接到一个或多个任务，并且任务可能与一个或多个 DWA 相关联，尽管有些任务没有任何关联的 DWA。

3.2 工资、就业和人口统计数据

我们从劳工统计局提供的 2020 年和 2021 年职业就业系列中获取就业和工资数据。该数据集包括职业名称、每个职业的工人人数和 2031 年的职业水平就业预测、进入职业所需的典型教育和获得职业能力所需的在职培训（BLS，2022）. 我们使用 BLS 推荐的 O*NET 人行横道（BLS，2023b）来链接 O*NET 任务和 DWA 数据集以及 BLS 劳动力人口统计数据（BLS，2023a），它来自当前人口调查（CPS）。这两个数据源均由美国政府收集，主要捕获非个体经营者、有记录且在所谓的正规经济中工作的工人。

3.3 曝光

我们根据曝光规则展示我们的结果，我们在其中定义**接触**作为衡量访问 LLM 或 LLM 支持的系统是否会减少人类执行特定 DWA 或完成任务所需时间至少 50% 的指标。尽管 GPT-4 具有视觉功能 OpenAI (2023b) 和“LLM”通常用于指代更广泛的模态，但视觉和图像功能仅

任务编号	职业名称	DWA	任务描述
14675	计算机系统 工程师/建筑师	监控计算机系统性能以确保正常运行。	监控系统运行以检测潜在问题。
18310	急症护理护士	操作诊断或治疗医疗仪器或设备。准备医疗用品或设备以备使用。	设置、操作或监控侵入性设备和装置，例如结肠造口术或气管切开术设备、机械呼吸机、导管、胃肠管和中心导管。
4668.0	赌笼 工人	执行销售或其他金融交易。	为顾客进行现金支票和处理信用卡预付款。
15709	网上商户	执行销售或其他金融交易。	发送已完成交易和装运的电子邮件确认。
6529	幼儿园 教师，除了 特殊教育	–	让家长志愿者和年长的学生参与儿童活动，以促进他们参与专注、复杂的游戏。
6568	小学 教师，除了 特殊教育	–	让家长志愿者和年长的学生参与儿童活动，以促进他们参与专注、复杂的游戏。

表 1：来自 O*NET 数据库的职业、任务和详细工作活动示例。我们发现仅对活动进行汇总是不精确的，事实证明，我们希望 Gambling Cage Workers 亲自完成给定的 DWA，使用一些体力，而我们希望在线商家仅使用计算机完成相同的活动.

包含在我们对 LLM 支持的软件的定义中。我们在下面提供了我们的规则摘要，而完整的规则可以在 A.1 中找到。当我们有 DWA 的标签时，我们首先将它们聚合到任务级别，然后再聚合到职业级别。

曝光量规摘要

无曝光 (E0) 如果：

- 使用所描述的 LLM 不会导致完成活动或任务所需的时间减少或减少最少，同时保持同等质量[A]或者
- 使用描述的 LLM 会导致活动/任务输出的质量下降。

直接暴露 (E1) 如果：

- 通过 ChatGPT 或 OpenAI playground 使用所描述的 LLM 可以将完成 DWA 或任务所需的时间减少至少一半 (50%)。

LLM+ 暴露 (E2) 如果：

- 单独访问所描述的 LLM 不会将完成活动/任务所需的时间减少至少一半，但是
- 可以在 LLM 之上开发额外的软件，这可以将完成特定活动/任务所需的时间减少至少一半。在这些系统中，我们计算了对图像生成系统的访问。[b]

[A]同等质量意味着第三方（通常是输出的接收者）不会注意到或关心 LLM 援助。

[b]在实践中，正如附录 A.1 中的完整规则所示，我们将对图像功能的访问单独分类 (E3) 以方便注释，尽管我们将 E2 和 E3 结合用于所有分析。

我们将曝光阈值设置为完成特定 DWA 或任务所需时间可能减少 50%，同时保持一致的质量。我们预计采用率将是最高和最直接的

适用于实现生产率显着提高的应用程序。尽管这个阈值有些随意，但选择它是为了便于注释者解释。此外，无论选择的阈值如何，我们猜测现实世界中任务时间的减少可能会略低于或显着低于我们的估计，因此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 LLM 或 LLM 支持的软件是否可以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，专栏 1；人类		E1	80.8%	0.223
		E1 + .5*E2	65.6%	0.591
	Z	E1 + E2	82.1%	0.654
GPT-4，专栏 2；人类		E1	81.8%	0.221
		E1 + .5*E2	65.6%	0.538
	Z	E1 + E2	79.5%	0.589
GPT-4，专栏 1；GPT-4，专栏 2		E1	91.1%	0.611
		E1 + .5*E2	76.0%	0.705
	Z	E1 + E2	82.4%	0.680

表 2：协议和 Pearson 相关分数的模型和人类比较。一致性分数是通过查看两组对注释（例如 E0、E1 或 E2）达成一致的频率来确定的。在本文中，我们使用 GPT-4，Rubric 1。

然后，我们使用曝光规则收集了人类和 GPT-4 生成的注释，这是本文大部分分析的基础。

- *人类评级：*我们通过将规则应用于每个 O*NET 详细工作人员活动 (DWA) 和所有 O*NET 任务的子集来获得人工注释，然后汇总这些 DWA 和任务分数5个在任务和职业层面。作为 OpenAI 对齐工作的一部分，作者亲自标记了大量任务和 DWA 样本，并招募了经验丰富的人工注释者，他们审查了 GPT-3、GPT-3.5 和 GPT-4 输出（Ouyang 等人，2022 年）。

- *GPT-4 评级：*我们管理了与早期版本的 GPT-4（OpenAI，2023b）类似的规则，但针对的是所有任务/职业对，而不是 DWA。我们对评分细则（在本例中用作模型的“提示”）进行了细微修改，以增强与一组人工标签的一致性。完全一致率见表 2。

我们为我们感兴趣的因变量构建了三个主要措施：（i），对应于上面暴露量规中的 E1，预计代表职业中暴露任务比例的下限，（ii），它是 E1 和 0.5*E2 的总和，其中 E2 的 0.5 权重旨在说明通过补充工具和应用程序部署技术需要额外投资时的风险，以及 (iii)*Z*，E1 和 E2 的总和，暴露的上限，提供对 LLLM 和 LLM 支持的软件的最大暴露的评估。我们在表 2 中总结了注释组和度量之间的一致性。对于分析的其余部分，如果未指定，读者可能会假设我们指的是 暴露——这意味着通过 ChatGPT 或 OpenAI Playground 等工具直接暴露的所有任务被认为是需要一些互补创新的任务的两倍。

5个作者注释了明显需要高度体力或手动灵巧性的 DWA，合同注释者标记了剩余的活动，以及任务的子集，包括那些没有相关 DWA 的任务和那些在聚合后没有明确任务级注释的任务DWA 注释。

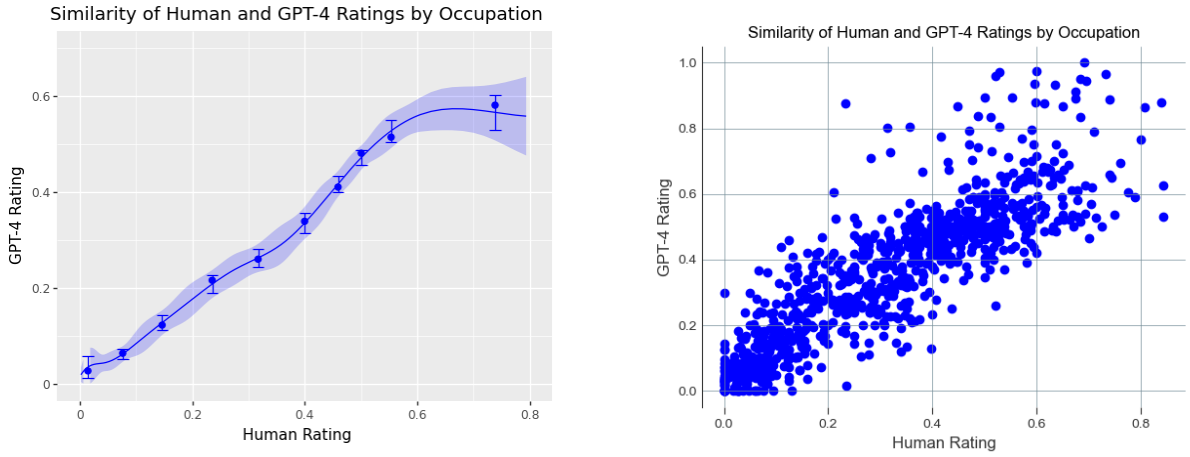


图 2：人类评分者（x 轴）和 GPT-4 评分者（y 轴）显示了按职业划分的 LLM 暴露程度高度一致。接近以下最高水平的暴露 聚合职业接触分数的方法，GPT-4 评级往往低于人类评级。我们展示了原始散点图和 binscatter。在接近暴露等级的顶端时，人类平均更有可能将职业评为暴露等级。

3.4 我们方法的局限性

3.4.1 人为主观判断

我们方法的一个基本限制在于标签的主观性。在我们的研究中，我们聘请了熟悉 LLM 能力的注释者。然而，这个群体在职业上并不多样化，这可能导致对法学硕士在不熟悉的职业中执行任务的可靠性和有效性做出有偏见的判断。我们承认，要为职业中的每项任务获得高质量的标签，就需要从事这些职业的工人，或者至少对这些职业中的各种任务有深入的了解。这代表了未来验证这些结果的重要工作领域。

3.4.2 用 GPT-4 测量 LLM

最近的研究表明，GPT-4 是一种有效的鉴别器，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 任务分类的结果对标题措辞的变化、提示的顺序和组成、标题中特定示例的存在与否、提供的详细程度以及关键术语的定义很敏感。根据在小型验证集中观察到的结果对提示进行迭代，可以增强模型输出与标题意图之间的一致性。因此，呈现给人类的标题与用于 GPT-4 的标题之间存在细微差别。故意做出此决定是为了在不过度影响人工注释者的情况下将模型引导至合理的标签。因此，我们使用多个注释源，但没有一个应该被认为是相对于其他注释的确定的基本事实。在此分析中，我们将人工注释者的结果作为我们的主要结果。在为 LLM 分类制定有效的规则方面仍有可能进一步改进和创新。尽管如此，我们仍然观察到人类评级和 GPT-4 评级在职业级别上关于 LLM 系统整体暴露的高度一致（参见表 2，图 2）。

3.4.3 其他弱点

- **基于任务的框架的有效性。**目前尚不清楚职业在多大程度上可以完全分解为任务，以及这种方法是否系统地忽略了某些类别的技能或任务，这些技能或任务是胜任工作所默认需要的。此外，任务可以由子任务组成，其中一些任务比其他任务更容易自动化。一些任务可能作为其他任务的先行任务，这样下游任务的完成就依赖于先行任务。如果确实，基于任务的分解不能有效地表示职业中的大部分工作是如何进行的，那么我们的暴露分析将在很大程度上失效。

- **缺乏专业知识和任务解释。**在标记过程中，人类注释者大多不知道映射到每个 DWA 的特定职业。这导致聚合任务和职业的逻辑不清晰，以及标签中的一些明显差异，如表 1 所示。我们试验了各种聚合方法，发现即使使用最大匹配方法（采用匹配的人类 <> 模型标签如果存在的话），协议保持相对一致。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。

- **具有前瞻性且可能会发生变化，并提供一些早期证据。**准确预测未来的 LLM 应用程序仍然是一项重大挑战，即使对于专家也是如此（OpenAI，2023b）。新出现的能力的发现、人类感知偏差的变化以及技术发展的转变都会影响关于 LLM 对工人任务的潜在影响以及 LLM 支持的软件开发的预测的准确性和可靠性。我们的预测本质上是前瞻性的，并基于当前的趋势、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，一些对于 LLM 或 LLM 支持的软件来说似乎不太可能影响到今天的任务可能会随着新模型功能的引入而改变。相反，暴露的任务可能会面临限制语言模型应用程序的不可预见的挑战。

- **分歧的来源。**虽然我们没有严格检查分歧的来源，但我们发现了人类和模型在评估中容易“卡住”的几个地方：
  - 虽然 LLM 理论上可以帮助或完成任务，但采用它来完成任务或活动需要多人改变他们的习惯或期望（例如会议、谈判），
  - 目前有一些法规或规范要求或建议人类监督、判断或同理心（例如做出决定、咨询）的任务或活动，以及
  - 已经存在可以合理自动化任务的技术的任务或活动（例如进行预订）。

4 个结果

通用技术相对较少，其特点是普遍性、随着时间的推移而改进，以及重要的共同发明和溢出效应的发展（Lipsey 等人，2005 年）。我们对 LLM 对劳动力市场的潜在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，法学硕士也可能影响这些方面。

在这个阶段，一些通用技术标准比其他标准更容易评估。我们在这个早期阶段的主要重点是检验 LLM 对经济具有普遍影响的假设，类似于 (Goldfarb et al., 2023) 采取的方法，他通过分析机器学习扩散

招聘信息以评估其作为通用技术的地位。我们不是一般地使用职位发布或研究机器学习，而是采用具有人类和 GPT-4 注释的任务评估方法。这种分析可能会揭示影响是否仅限于一组特定的类似任务或职业，或者它们是否会更广泛。

我们的研究结果表明，根据他们的任务级别能力，法学硕士有可能显着影响美国境内的各种职业。经济性，展示了通用技术的一个重要属性。在以下部分中，我们将讨论各种角色和工资结构的结果。有关美国经济中各行业相对风险的其他结果，请参见附录 D。

4.1 汇总统计

这些措施的汇总统计数据可在表 3 中找到。人类和 GPT-4 注释均表明平均职业水平 值介于 0.14 和 0.15 之间，这表明平均而言，职业中大约 15% 的任务直接接触 LLM。这个数字增加到 30% 以上 并超过 50%*Z*. 巧合的是，人类和 GPT-4 注释也将数据集中 15% 到 14% 的总任务标记为暴露给 LLM。基于值，我们估计 80% 的工人属于至少 10% 的任务暴露于 LLM 的职业，而 19% 的工人所处的职业超过一半的任务被标记为暴露。

我们使用 O*NET 的“重要性”分数运行了一组分析，但没有发现我们的发现有重大变化。尽管我们确实承认，不对任务对给定职业的相对重要性进行加权会产生一些奇怪的结果（例如，理发师的曝光率相当高）。

尽管任务受到影响的可能性很大，但 LLM 和 LLM 支持的软件必须纳入更广泛的系统才能充分发挥这种潜力。与通用技术一样，共同发明障碍最初可能会阻碍 GPT 快速扩散到经济应用中。此外，预测人类监督的需求具有挑战性，特别是对于模型能力等于或超过人类水平的任务。虽然对人工监督的要求最初可能会减慢这些系统在经济中扩散的速度，但 LLM 和 LLM 支持的系统的用户可能会随着时间的推移越来越熟悉该技术，特别是在了解何时以及如何相信它的输出。

职业水平暴露					
	人类		**GPT-4**		
	平均标准	平均标准	0.14		
	0.14	0.14	0.16	0.30	0.21
	0.34	0.22	0.46	0.30	0.55
Z	0.34				
任务级暴露					
	人类		**GPT-4**		
	平均标准	平均标准	0.15		
	0.36	0.14	0.35	0.31	0.37
	0.35	0.35	0.47	0.50	0.56
Z	0.50				

表 3：我们的人类和模型暴露数据的汇总统计。

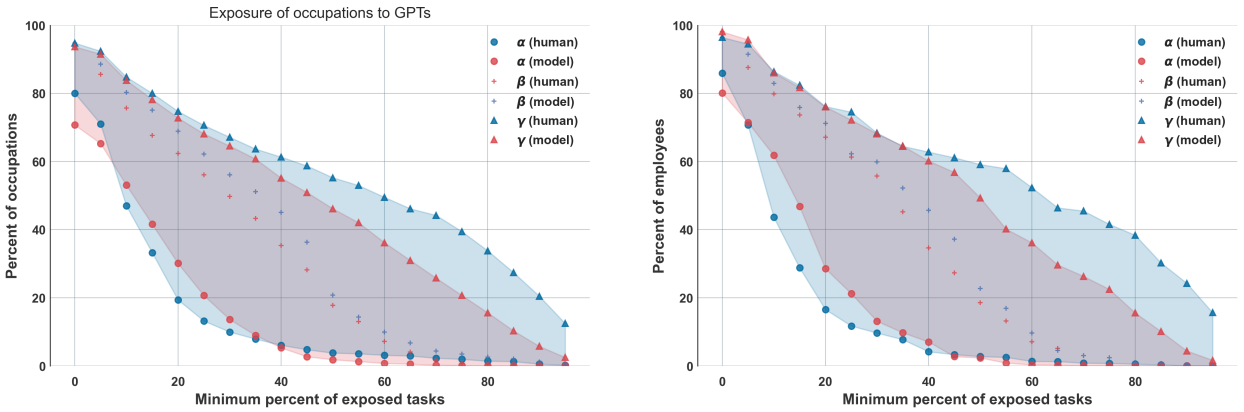


图 3：整个经济体的暴露强度，左侧显示受影响职业的百分比，右侧显示受影响工人的百分比。不同职业和不同工人的接触分布相似，这表明工人在职业中的集中度与职业接触 LLM 或 LLM 支持的软件并不高度相关。然而，我们确实希望它与为特定领域开发 LLM 支持的软件的投资高度相关。

4.2 工资和就业

在图 3 中，我们展示了整个经济体的风险敞口强度。第一个图显示职业方面的暴露，而第二个图显示工人总数方面的暴露。图表上的每个点代表 y 轴上具有暴露水平的工人（和职业）的估计百分比（，，和*Z*) 表示在 x 轴上。例如，人工注释者确定 2.4% 的工人是 50-暴露，18.6%是 50-暴露，49.6%是*Z*50 -exposed，其中 50% 的阈值来自 x 轴，工人的百分比来自图 2 右图中的 y 轴。在 x 轴上的任何给定点，工人之间的垂直距离 和*Z*表示除了直接暴露于 LLM 之外，归因于工具和应用程序的暴露可能性。工人和职业的暴露分布相似，这表明工人在职业中的专注度与 LLM 或 LLM 支持的软件的职业暴露没有很强的相关性。

如图 4 所示，在职业层面汇总，人类和 GPT-4 注释表现出定性的相似性并且倾向于相关。与 GPT-4 注释相比，人类注释估计高薪职业的曝光率略低。虽然有许多低薪职业具有高暴露度，而高薪职业具有低暴露度，但 binscatter plot 的总体趋势表明，较高的工资与 LLM 的暴露度增加有关。

LLM 的潜在风险似乎与当前的就业水平几乎没有关联。在图 4 中，人类和 GPT-4 的整体暴露评级都汇总到职业级别（y 轴），并与总就业（x 轴）的对数进行比较。这两个图都没有显示不同就业水平的 LLM 暴露存在显着差异。

4.3 技能重要性

在本节中，我们探讨了职业技能的重要性（如 O*NET 数据集中的注释）与我们的暴露度量之间的关系。首先，我们使用 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度量进行归一化，以提高结果的可理解性。接下来，我们对我们的暴露措施进行回归分析（，，*Z*) 来检查技能重要性和接触之间的关联强度。

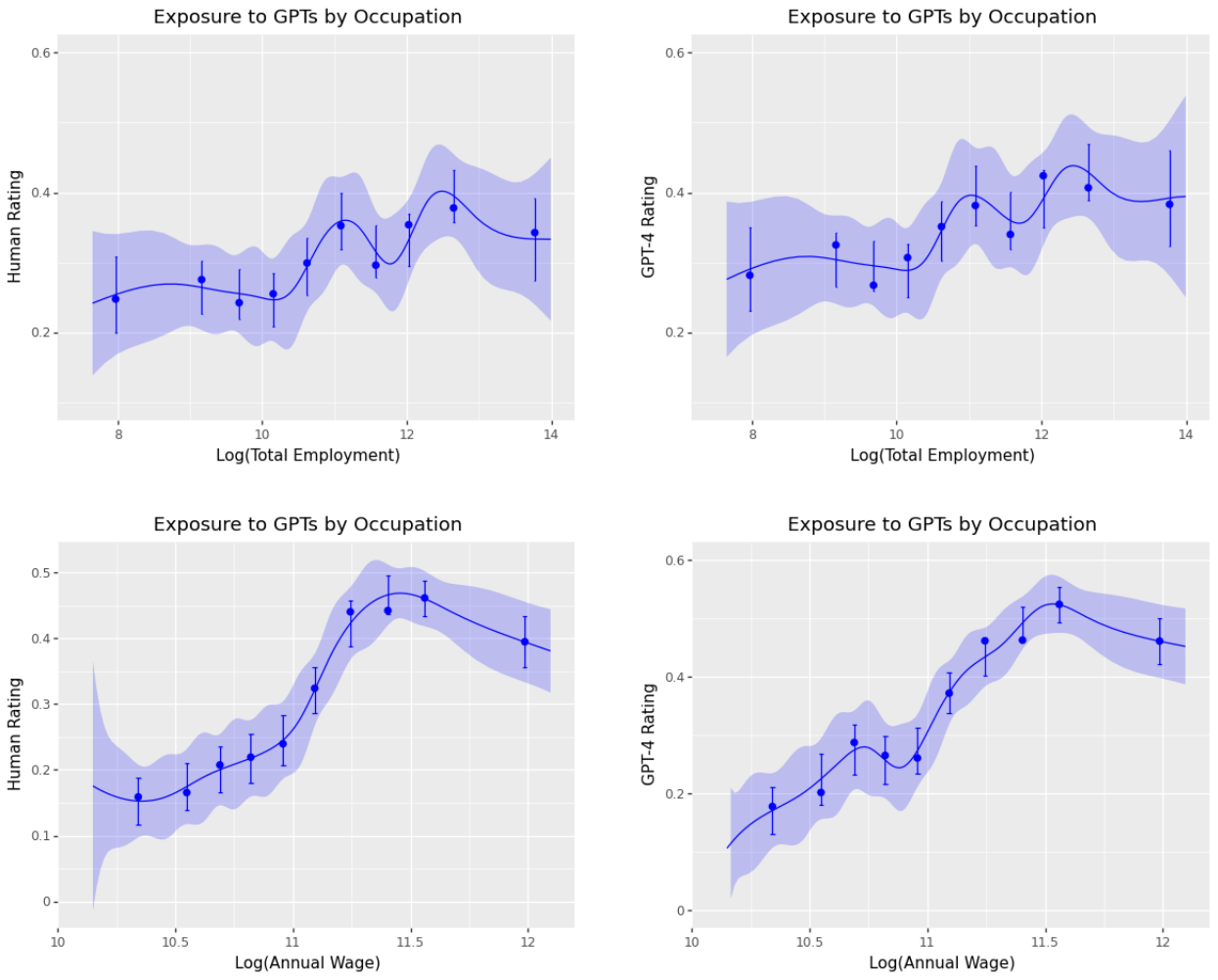


图 4：binscatter plots 描绘了各种职业中语言模型 (LLM) 的暴露，由人类评估者和 GPT-4 评估。这些图比较了对 LLM 和部分 LLM 支持的软件的暴露（) 在职业水平与职业中总就业人数的对数和职业年工资中位数的对数。虽然存在一些差异，但人类和 GPT-4 评估都表明，高薪职业往往更容易接触 LLM。此外，根据我们的标准，许多低薪职业表现出高曝光率。在计算平均暴露分数时，核心任务的权重是职业中补充任务的两倍。就业和工资数据来自 2021 年 5 月进行的 BLS-OES 调查。

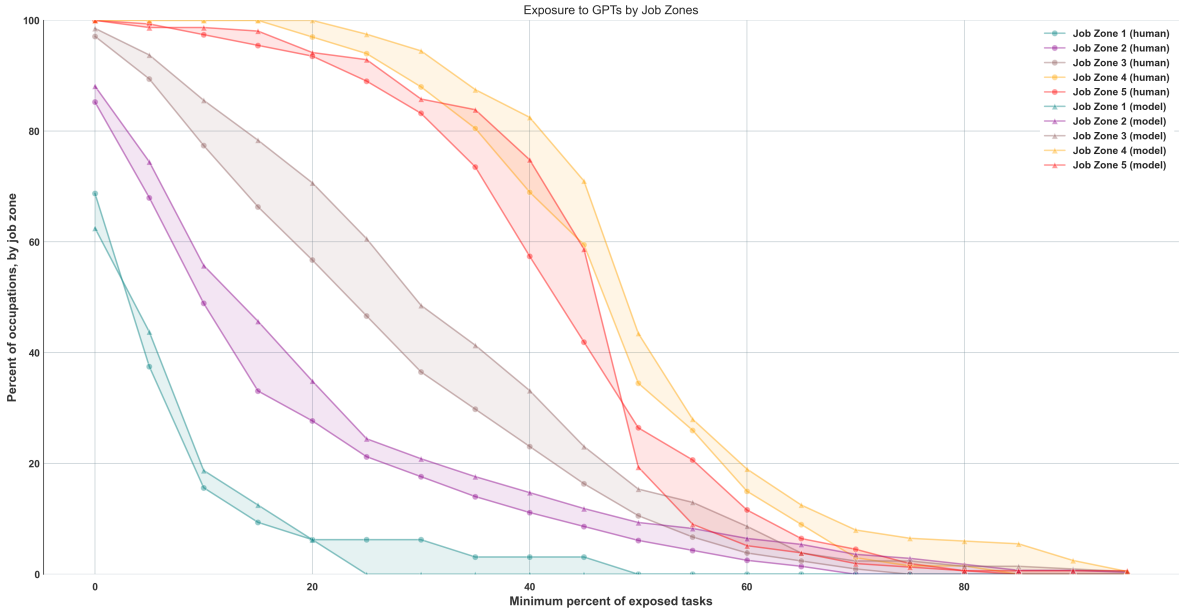


图 5： 五个 JobZones 中职业的暴露等级，这些职业是根据教育水平、经验和 从事这些工作所需的在职培训进行分类的相似职业组。

我们的研究结果表明**科学**和**批判性思维**技能与接触程度呈强烈负相关，这表明需要这些技能的职业不太可能受到当前 LLM 的影响。反过来，**编程**和**写作**技能与暴露程度呈正相关，这意味着涉及这些技能的职业更容易受到 LLM 的影响（详细结果见表 5）。

4.4 进入壁垒

接下来，我们检查进入壁垒，以更好地了解是否因工作类型而存在风险敞口差异。一种这样的代理是称为“工作区”的 O*NET 职业级别描述符。工作区将以下方面相似的职业分组：(a) 在职业中找到一份工作所需的教育水平，(b) 从事该工作所需的相关经验的数量，以及 (c) 工作的程度工作所需的工作培训。在 O*NET 数据库中，有 5 个工作区，工作区 1 需要最少的准备（3 个月），工作区 5 需要最广泛的准备，4 年或更长时间。我们观察到随着所需准备水平的增加，工作区的收入中位数单调增加，工作区 1 中的中位数工人收入为 30 美元,工作区 5 的 230 人和收入 80 美元的中等工人,980。

我们所有的措施（, , 和 *Z*) 显示相同的模式，即从工作区 1 到工作区 4 的暴露增加，并且在工作区 5 保持相似或减少。与图 3 类似，在图 5 中，我们绘制了每个阈值的工人百分比接触。我们发现，平均而言，从事 50% 以上职业的工人所占百分比 在工作区 1 到 5 的曝光有 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 入学所需的典型教育

由于包含在工作区中既考虑了所需的教育——这本身就是技能获取的代表——也考虑了所需的准备，因此我们寻求数据来理清这些变量。我们使用来自劳工统计局的职业数据的两个变量：“TypicalEducationNeededforEntry”和“On-the-job”

获得职业能力所需的培训。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。我们缺乏3,504,000 名工人的教育和在职培训要求数据，他们是因此排除在汇总表之外。

我们的分析表明，与没有正规教育证书的人相比，持有学士、硕士学位和专业学位的人更容易接触到 LLM 和 LLM 支持的软件（见表 7）。有趣的是，我们还发现受过一些大学教育但没有学位的人对 LLM 和 LLM 支持的软件表现出较高的接触程度。在检查显示进入壁垒的表格后，我们发现接触最少的工作需要最多的培训，一旦获得能力，可能会提供较低的回报（就收入中位数而言）。相反，不需要在职培训或只需要实习/居住的工作似乎会产生更高的收入，但更容易获得法学硕士学位。

团体	接触率最高的职业	% 接触
人类	口译员和笔译员调查研究	76.5
	人员	75.0
	诗人、作词家和创意作家 动物科学	68.8
	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
	口译员和笔译员公共关系	82.4
	专家动物科学家	80.6
		77.8
人类*Z*	数学家	100.0
	报税员	100.0
	金融量化分析师作家和作者	100.0
		100.0
	网络和数字界面设计师	100.0
	人类将 15 种职业标记为 “完全暴露”。	
模型	数学家	100.0
	通讯员	95.2
	区块链工程师	94.1
	法庭记者和同步字幕员校对员和复制标记	92.9
		90.9
模型	数学家	100.0
	区块链工程师	97.1
	法庭记者和同步字幕员校对员和复制标记	96.4
		95.5
	通讯员	95.2
模型*Z*	会计师和审计师	100.0
	新闻分析师、记者和记者法律秘书和行政助	100.0
	理临床数据经理	100.0
		100.0
	气候变化政策分析师	100.0
	该模型将 86 个职业标记为 “完全暴露”。	
最高方差	搜索营销策略师平面设计师	14.5
		13.4
	投资基金经理 财务经理	13.0
		13.0
	保险估价师，汽车损坏	12.6

表 4：根据每次测量，暴露程度最高的职业。最后一行列出了最高的职业 2个值，表明他们在曝光分数上的变异性最大。暴露百分比表示暴露于 GPT 的职业任务的份额（) 或 GPT 驱动的软件 (和*Z*)，其中暴露被定义为将完成任务所需的时间减少至少 50%（参见暴露规则 A.1）。因此，此表中列出的职业是我们估计 GPT 和 GPT 支持的软件能够为员工节省大量时间来完成大部分任务的职业，但这并不一定表明他们的任务可以完全完成由这些技术自动化。

基本技能			*Z*
	(标准错误)	(标准错误)	(标准错误)
	所有技能重要性分数都标准化为 0 到 1 之间。		
持续的	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)
请讲	- 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)
监控	- 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以及编程技能的职业级别、人工注释暴露于 GPT 对技能重要性的回归。技能的描述可以在附录 B 中找到。

工作区	准备必需的	教育必需的	示例职业	中位数收入	总公司 (000s)	H 米	H 米	H 米
								Z *Z*
1个	没有或很少(0-3个月)	中学文凭或GED（可选）	食品准备工人、洗碗机、地板砂光机	30,230 美元	13,100	0.03 0.04	0.06 0.06	0.09 0.08
2个	一些（3-12个月）	中学文凭	勤务员、客户服务代表、柜员	38,215 美元	73,962	0.07 0.12	0.16 0.20	0.24 0.27
3个	中（1-2年）	职业学校，在职培训，或同事的程度	电工、理发师、医疗助理	54,815 美元	37,881	0.11 0.14	0.26 0.32	0.41 0.51
4个	大量（2-4 岁）	学士学位	数据库管理员、平面设计师、成本估算员	77,345 美元	56,833	0.23 0.18	0.47 0.51	0.71 0.85
5个	广泛的（4+年）	硕士或更高	药剂师、律师、天文学家	81,980 美元	21,221	0.23 0.13	0.43 0.45	0.63 0.76

表 6：按工作区域划分的 GPT 平均暴露程度。对于每个工作区，我们还提供了每个构成职业的年收入中位数（以美元为单位），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	总 Emp (000s)	H	米	H	米	H*Z*	米*Z*
没有任何	77,440 美元	90,776	0.20	0.16	0.42	0.46	0.63	0.76
学徒制	55,995 美元	3,066	0.01	0.02	0.04	0.06	0.07	0.10
实习/居住	$77,110	3,063	0.16	0.06	0.36	0.38	0.55	0.71
短期在职培训 中期在职培训 长期在职培训	33,370 美元	66,234	0.11	0.15	0.21	0.25	0.32	0.34
	46,880 美元	31,285	0.09	0.12	0.21	0.25	0.32	0.38
	48,925 美元	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，按获得工作能力所需的在职培训水平分组。除了曝光分数，我们还显示了每个职业的年收入中位数，以及每个组中的工人总数，以千为单位。

5 措施的验证

5.1 与早期努力的比较

本文旨在建立在之前的一些实证研究的基础上，这些研究考察了职业暴露于人工智能和/或自动化进步的情况。以前的研究使用了多种方法，包括：

- 使用像 O*NET 这样的职业分类法来描述哪些职业具有常规与非常规以及手动与认知任务内容（Autor 等人，2003 年；Acemoglu 和 Autor，2011a）。
- 将任务的文本描述映射到专利技术进步的描述。（Kogan 等人，2021 年；Webb，2020 年）
- 将人工智能系统的能力与职业能力联系起来，并将暴露估计汇总到需要这些能力的职业。（Felten 等人，2018 年、2023 年）
- 通过从认知科学文献中提取的一组 14 种认知能力，将 AI 任务基准评估（ImageNet、Robocup 等）的结果映射到 59 个工人任务。（托兰等人，2021 年）
- 专家对一组专家高度信任的 O*NET 职业的自动化潜力进行专家标记，并结合概率分类器来估计其余 O*NET 职业的自动化潜力。（弗雷和奥斯本，2017 年）
- 制定一个标准来评估工人在经济中完成的活动的“机器学习的适用性”（SML）（Brynjolfsson 和 Mitchell，2017 年；Brynjolfsson 等人，2018 年，2023 年）。

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数据。

本文的方法主要建立在 SML 方法的基础上，通过开发一个规则来评估 O*NET 数据库中报告的 LLM 功能和工作任务之间的重叠。表 9 显示了我们新的 LLM 暴露测量对职业水平暴露测量的 OLS 回归结果（Felten 等人，2018 年）（表中的“AI 职业暴露评分”），（Frey 和 Osborne，2017 年）（Frey & Osborne Automation），（Webb，2020）中所有三种技术的得分，（Acemoglu 和 Autor，2011a）中的标准化常规手动和认知得分，以及（Brynjolfsson 等，2018、2023）（SML）。我们还使用来自最新 BLS 职业就业调查的年化职业工资作为对照。

GPT-4 曝光评级 1 对应于我们由 GPT-4 评估的整体曝光量规，其中完全曝光潜力编码为 1，没有曝光潜力编码为 0，部分曝光（我们的标签方案中的 E2）编码为 0.5 . GPT-4 曝光等级 2 的总体曝光评分相似，但提示略有不同。两个提示的结果非常相似。人类暴露评级代表与 GPT-4 暴露评级 1 中相同的标准，但由人类评分，如本文前面部分所述。这些结果对应于 上面给出的一组统计数据。

每种测量类型的结果都是一致的。我们发现我们的 LLM 暴露度量与之前针对软件和 AI 的度量之间通常存在正相关和统计显着相关性。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显着且正相关的关系，证明了采用相似方法的两项研究之间存在一定程度的凝聚力。Webb 软件和基于 AI 专利的测量、SML 和归一化（贬低并除以标准差）常规认知分数都与我们的一些测量呈正相关。

	最小值	25%	中位数	75th Perc	最大限度	平均标准。	开发。	数数
GPT-4 暴露等级 1 GPT-4	0.00	0.13	0.34	0.50	1.00	0.33	0.22	750
暴露等级 2 人类暴露等级	0.00	0.09	0.24	0.40	0.98	0.26	0.20	750
	0.00	0.09	0.29	0.47	0.84	0.29	0.21	750
软件（网络）	1.00	25.00	50.00	75.00	100.00	50.69	30.05	750
机器人（韦伯）	1.00	22.00	52.00	69.00	100.00	48.61	28.61	750
人工智能（韦伯）	1.00	28.00	55.00	82.00	100.00	54.53	29.65	750
Suitability for Machine	2.60	2.84	2.95	3.12	3.55	2.99	0.18	750
Learning Normalized Routine	- 3.05	- 0.46	0.10	0.63	3.42	0.07	0.86	750
认知归一化例程手册	- 1.81	- 0.81	- 0.11	0.73	2.96	0.05	1.01	750
AI 职业暴露分数 Frey	1.42	3.09	3.56	4.04	6.54	3.56	0.70	750
&Osborne 自动化日志平均。薪	0.00	0.07	0.59	0.88	0.99	0.50	0.38	681
水	10月13日	10.67	11.00	11.34	12.65	11.02	0.45	749

表 8：衡量职业接触 AI 和自动化的一系列先前工作的汇总统计数据。我们还包括了这项工作中新出现的测量的汇总统计数据。我们包括来自 (Webb, 2020) 的所有测量值、来自 (Acemoglu 和 Autor, 2011a) 的标准化常规认知和手动评分（由于职业群体的不完美匹配，平均值可能略微偏离 0）、来自（Brynjolfsson 和 Mitchell）的机器学习适用性，2017 年；Brynjolfsson 等人，2018 年，2023 年），来自（Felten 等人，2018 年）的 AI 职业暴露，以及来自（Frey 和 Osborne，2017 年）的自动化暴露。我们包括了尽可能多的职业，但由于 O*NET 分类法随着这些措施的制定而发生了变化，因此最新版本的 O*NET 6 位数职业中可能缺少一些角色。

软件、SML 和常规认知分数都显示出与 LLM 接触分数在 1% 水平上的正向和统计显着关联。(Webb, 2020) 的 AI 分数系数也为正且在 5% 的水平上具有统计显着性，但我们在第 3 列和第 4 列中对 LLM 总体接触的次要提示并未显示出具有统计显着性的关系。在大多数情况下，AI 职业暴露评分与我们的暴露指标无关。Webb 的机器人曝光分数、常规手动任务内容和整体自动化指标（Frey 和 Osborne，2017 年）都与我们的主要 GPT-4 和人工评估的整体曝光评级呈负相关，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。

与 (Felten et al., 2018) 和 (Frey and Osborne, 2017) 的低相关性可能可以用方法的差异来解释。将 AI 能力与工人能力联系起来或直接根据职业特征对暴露进行评分，而不是从 DWA 或任务级评分（如 SML 论文和我们自己的）汇总到职业，提供了对内容略有不同的看法职业。

在所有回归中， 2个介于 60.7%（第 3 列）和 72.8%（第 5 列）之间。这表明，与其他测量相比，我们明确关注 LLM 能力的测量具有 28% 到 40% 的无法解释的方差。特别是在与 AI 相关的曝光分数的情况下，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 或 LLM 支持的软件的未来进展的信息有限。我们希望我们对未来机器学习技术的理解同样不完美地体现在我们今天的标准中。

6 讨论

6.1 GPT 作为通用技术

在本文的前面，我们讨论了 LLM 可以归类为通用技术的可能性。这种分类要求 LLM 满足三个核心标准：随着时间的推移而改进，贯穿始终

	GPT-4 曝光等级 1		GPT-4 曝光等级 2		人体暴露等级	
	(1)	(2)	(3)	(4)	(5)	(6)
软件（网络）	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（韦伯）	−0.00378***	−0.00405***	−0.00377***	−0.00399***	−0.00371***	−0.00383***
	(0.00032)	(0.00031)	(0.00034)	(0.00033)	(0.00029)	(0.00028)
人工智能（韦伯）	0.00080***	0.00090***	0.00036	0.00045	0.00067**	0.00071**
	(0.00030)	(0.00029)	(0.00030)	(0.00030)	(0.00030)	(0.00030)
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)
标准化常规认知	0.06601***	0.06868***	0.04743***	0.05015***	0.03568***	0.03659***
	(0.00886)	(0.00894)	(0.00872)	(0.00879)	(0.00671)	(0.00669)
规范化例行手册	−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 职业暴露分数	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)
日志平均。薪水	0.05804***		0.04863***		0.02531	
	(0.01870)		(0.01860)		(0.01727)	
持续的	−1.12937***	−0.45743***	−0.96117***	−0.39935***	−0.47078*	−0.17706
	(0.26859)	(0.15327)	(0.26365)	(0.15017)	(0.24684)	(0.13256)
否	680.00000	681.00000	680.00000	681.00000	680.00000	681.00000
2个	0.68741	0.68212	0.60737	0.60198	0.71213	0.71126

表 9：LLM 暴露分数回归 AI 和自动化职业暴露的先前测量。我们还包括 2021 年 5 月 BLS-OES 调查的年化工资。每项指标都保持在其原始规模，但来自（Acemoglu 和 Autor，2011a）的常规认知和常规手动评分除外。这两个分数被标准化为均值 0 和方差 1。通常我们发现与以前的努力有很强的正相关，尽管我们的新措施仍然可以解释大的剩余方差。第 1 列和第 2 列基于我们的主要 来自 GPT-4 评级的曝光度量。第 3 列和第 4 列基于类似的略有不同的曝光量规，该量规也由 GPT-4 评定为稳健性。第 5 列和第 6 列反映了人类对与第 1 列和第 2 列相同的评分标准的评分。

经济，以及催生互补创新的能力（Lipsey 等人，2005 年）。来自 AI 和机器学习文献的证据彻底证明了 LLM 满足第一个标准——随着时间的推移，它们的能力正在提高，能够完成或帮助日益复杂的任务和用例集（见 2.1）。本文提供了支持后两个标准的证据，发现 LLM 本身可以对整个经济产生普遍影响，并且 LLM 实现的互补创新——特别是通过软件和数字工具——可以广泛应用于经济活动。

图 3 展示了构建在 LLM 之上的补充软件的潜在经济影响。取 y 轴（所有职业的份额）之间的差异和Z在 x 轴上的给定点（职业中暴露的任务的份额）给出了职业内总的暴露潜力，这些潜在暴露归因于工具和软件，而不是 LLM 本身的直接暴露。所有任务之间的均值差异 和Z使用 GPT-4 注释为 0.42，使用人工注释为 0.32（见图 3），这表明 LLM 支持的软件对任务暴露的平均影响可能是 LLM 自身的平均暴露的两倍多（平均Z基于人工注释和 GPT-4 注释的 0.14）。虽然我们的研究结果表明这些开箱即用的模型与有意义的工作人员和任务份额相关，但它们也表明它们产生的软件创新可以产生更广泛的影响。

一项技术的普及程度的一个组成部分是其被企业和用户采用的程度。本文没有系统地分析这些模型的采用情况，但是，早期的定性证据表明，LLM 的采用和使用正变得越来越普遍。在 LLM 之上相对简单的 UI 改进的力量在 ChatGPT 的推出中显而易见——其中底层语言模型的版本以前可以通过 API 获得，但在 ChatGPT 界面发布后使用量猛增。（Chow，2023 年；OpenAI，2022 年）此版本发布后，多项 商业调查表明，在过去几个月中，公司和员工对 LLM 的采用有所增加。（康斯坦茨，2023 年；ResumeBuilder.com，2023 年）

这些模型的广泛采用需要解决现有的瓶颈。它们效用的一个关键决定因素是人类对它们的信任程度以及人类如何适应他们的习惯。例如，在法律专业中，模型的有用性取决于法律专业人士是否可以在不验证原始文件或进行独立研究的情况下信任模型输出。技术的成本和灵活性、工人和公司的偏好以及激励措施也会显着影响基于 LLM 构建的工具的采用。通过这种方式，采用可能会受到与 LLM 相关的一些道德和安全风险的进展所驱动：偏见、捏造事实和偏差，举几个例子 OpenAI (2023a)。此外，由于数据可用性等因素，不同经济部门对 LLM 的采用会有所不同，监管环境，以及权力和利益的分配。因此，全面了解工人和公司对 LLM 的采用和使用需要对这些错综复杂的问题进行更深入的探索。

一种可能性是，对于大多数任务而言，节省时间和无缝应用比提高质量更重要。另一个是最初的重点将放在增强上，然后是自动化（Huang 和 Rust，2018）。这可能形成的一种方式是通过增强阶段，在该阶段工作首先变得更加不稳定（例如，作家成为自由职业者），然后再过渡到完全自动化。

6.2 对美国公共政策的影响

包括法学硕士在内的自动化技术的引入以前与经济差距加剧和劳动力中断有关，这可能会对下游产生不利影响（Acemoglu 和 Restrepo，2022a；Acemoglu，2002；Moll 等，2021；Klinova 和 Korinek , 2021 年；Weidinger 等人，2021 年、2022 年）。我们对美国工人暴露情况的研究结果强调，需要社会和政策准备应对 LLM 及其产生的互补技术造成的潜在经济破坏。虽然建议具体的政策处方不在本文的范围之内

为了顺利过渡到越来越广泛采用 LLM 的经济，之前的工作（Autor 等人，2022b）已经阐明了美国与教育、工人培训、安全网计划改革等相关政策的几个重要方向。

6.3 局限性和未来工作

除了上面讨论的那些之外，我们还强调了这项工作的一些特殊局限性，这些局限性值得进一步调查。首先，我们对美国的关注限制了我们的研究结果对其他国家的普遍适用性，在这些国家，生成模型的采用和影响可能因工业组织、技术基础设施、监管框架、语言多样性和文化背景等因素而有所不同。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，以便其他研究人员可以在此基础上进行研究。

随后的研究工作应考虑另外两项研究：一项探索不同部门和职业的 LLM 采用模式，另一项研究与超出我们曝光分数范围的工人活动相关的最先进模型的实际能力和局限性。例如，尽管最近 GPT-4 在多模态能力方面取得了进展，但我们没有考虑视觉能力 对直接 LLM 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。此外，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、开放式和特定领域的任务中。

7 结论

总之，本研究考察了 LLM 对美国经济中各种职业和行业的潜在影响。通过应用新的规则来理解 LLM 能力及其对工作的潜在影响，我们观察到大多数职业都表现出一定程度的 LLM 暴露，而高薪职业通常会呈现更多暴露程度高的任务。我们的分析表明，在考虑当前模型功能和预期的 LLM 支持的软件时，大约 19% 的工作将至少 50% 的任务暴露给 LLM。

我们的研究旨在突出 LLM 的通用潜力及其对美国工人的可能影响。以前的文献展示了迄今为止 LLM 令人印象深刻的改进（见 2.1）。我们的研究结果证实了这样一个假设，即这些技术可以对美国的广泛职业产生普遍影响，并且 LLM 支持的额外进步（主要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 LLM 使人类劳动更有效率的技术能力似乎很明显，但重要的是要认识到社会、经济、监管和其他因素会影响实际的劳动生产率结果。随着能力的不断发展，LLM 对经济的影响可能会持续存在并增加，

需要进一步研究来探索 LLM 进步的更广泛影响，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成果。通过寻求了解 LLM 对劳动力的能力和潜在影响，政策制定者和利益相关者可以做出更明智的决策，以驾驭 AI 的复杂格局及其在塑造未来工作中的作用。

7.1 LLM 结论（GPT-4 的版本）

生成式预训练变形金刚 (GPT) 产生深刻的转变，获得潜在的技术增长，渗透任务，极大地影响专业。这项研究探讨了 GPT 的潜在轨迹，提出了一个开创性的标准来衡量任务的 GPT 风险，特别是在美国劳动力市场。

7.2 LLM 结论（作者增补版）

生成式预训练变压器 (GPT) 产生深刻的转变，获得潜在的技术增长，渗透任务，破坏专业管理。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，概括过去的今天。

致谢

感谢帮助我们注释任务暴露的注释者组，包括 Muhammad Ahmed Saeed、Bongane Zitha、MerveÖzen Şenen、JJ 和 Peter Hoeschele。我们还感谢 Lauryn Fuld、AshleyGlat、Michael Lampe 和 Julia Susser 提供的出色研究协助。我们感谢 Miles Brundage 对本文的重要反馈。

我们感谢 Todor Markov 和 Vik Goel 建立了我们用来使用 GPT-4 运行我们的规则的基础设施。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 JeffWu 在项目的各个阶段提供反馈和编辑。

LLM助学声明

GPT-4 和 ChatGPT 在该项目中用于编写、编码和格式化辅助。

A　专栏

A.1　接触

E 曝光量规

考虑最强大的 OpenAI 大型语言模型 (LLM)。该模型可以完成许多任务，这些任务可以表述为具有文本输入和文本输出，其中输入的上下文可以用 2000 个单词捕获。该模型也无法绘制最新的事实（<1 年前的事实），除非它们在输入中被捕获。

假设您是一名工作人员，在您的角色中具有平均水平的专业知识，试图完成给定的任务。您可以访问 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。您还可以使用通过笔记本电脑访问的任何常用技术工具（例如麦克风、扬声器等）。您无权使用任何其他物理工具或材料。

请根据下面的标题标记给定的任务。
同等质量意味着审查工作的人无法判断工作是人类自己完成的还是在 LLM 的帮助下完成的。

如果您不确定如何判断一项任务所花费的时间，请考虑所描述的工具是否公开了与该任务关联的大部分子任务。

E1 – 直接曝光

如果仅通过 ChatGPT 或 OpenAI playground 等接口直接访问 LLM，则将任务标记为 E1 可以将以同等质量完成任务所需的时间减少至少一半。这包括可以简化为的任务： - 根据复杂的指令编写和转换文本和代码， - 根据规范对现有文本或代码进行编辑， - 编写可以帮助执行过去手工完成的任务的代码， -在不同语言之间翻译文本， - 总结中等长度的文档，

- 提供有关文档的反馈， - 回答有关文档的问题， - 生成用户可能想问的有关文档的问题， - 为面试或评估编写问题， - 编写和回复电子邮件，包括涉及反驳信息或参与的电子邮件在谈判中（但仅限于通过书面通信进行的谈判）， - 保留书面数据的记录， - 根据常识准备培训材料，或 - 通过任何书面或口头媒介将任何信息告知任何人。

E2 – 由 LLM 驱动的应用程序曝光

如果单独访问 LLM，则将任务标记为 E2 可能不会将完成任务所需的时间减少至少一半，但很容易想象可以在 LLM 之上开发的其他软件可以减少所需的时间完成任务一半。该软件可能包括以下功能： - 总结超过 2000 字的文档并回答有关这些文档的问题， - 从 Internet 检索最新的事实并将这些事实与 LLM 功能结合使用，

- 搜索组织的现有知识、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， - 分析书面信息以告知决策， - 根据高度专业化的知识准备培训材料， - 就问题提供建议，以及 - 维护复杂的数据库。

E3 – 给定图像能力的曝光

假设您可以访问 LLM 和一个可以查看、描述和创建图像的系统，以及任何由 LLM 提供支持的系统（上面 E2 中的系统）。该系统不能将视频作为输入，也不能产生视频作为输出。该系统无法从图像输入中准确地检索非常详细的信息，例如图像内尺寸的测量值。如果在访问 LLM 和这些图像功能的情况下完成任务所需的时间显着减少，则将任务标记为 E3： - 从 PDF 中读取文本， - 扫描图像，或 - 根据说明创建或编辑数字图像。

图像可以是逼真的，但不应该是详细的。该模型可以识别图像中的对象，但不能识别这些选项之间的关系。

E0 – 无曝光

如果以上都不是，则将任务标记为 E0 明显减少了经验丰富的工人高质量完成任务所需的时间至少一半。一些例子： - 如果一项任务需要高度的人际互动（例如，面对面的演示），那么它应该被归类为 E0。- 如果一项任务需要精确测量，那么它应该被归类为 E0。- 如果一项任务需要详细审查视觉效果，则应将其归类为 E0。- 如果一项任务需要用手或步行，则应将其归类为 E0。- 建立在 LLM 之上的工具不能做出任何可能影响人类生计的决定（例如招聘、评分等）。如果任务的任何部分涉及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），则应将其归类为 E0。法学硕士可以提出建议。- 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时间来完成任务，那么它应该被归类为 E0。- 法学硕士和建立在它之上的系统不能做任何法律上要求人类执行任务的事情。- 如果现有技术不是由常用的 LLM 提供支持并且可以完成任务，那么如果使用 LLM 或 LLM 支持的工具不会进一步减少完成任务的时间，则应将任务标记为 E0。- 法学硕士和建立在它之上的系统不能做任何法律上要求人类执行任务的事情。- 如果现有技术不是由常用的 LLM 提供支持并且可以完成任务，那么如果使用 LLM 或 LLM 支持的工具不会进一步减少完成任务的时间，则应将任务标记为 E0。- 法学硕士和建立在它之上的系统不能做任何法律上要求人类执行任务的事情。- 如果现有技术不是由常用的 LLM 提供支持并且可以完成任务，那么如果使用 LLM 或 LLM 支持的工具不会进一步减少完成任务的时间，则应将任务标记为 E0。

如有疑问，您应该默认为 E0。

#注解示例：

职业：检查员、测试员、分拣员、取样员和称重员 任务：调整、清洁或修理产品或加工设备，以纠正检查期间发现的缺陷。标签（E0/E1/E2/E3）：E0 说明：该模型无法获得任何实体，描述的任务（调整、清洁和维修设备）超过一半需要手或其他实体。

职业：计算机和信息研究科学家任务：应用理论知识和创新来创造或应用新技术，例如将计算机应用到新用途的适应原理。标签（E0/E1/E2/E3）：E1 解释：模型可以在训练过程中学习理论知识作为其一部分

一般知识库和适应原则可以在模型的文本输入中捕获。

活动：安排用餐预订。标签（E0/E1/E2/E3）：E2 解释：为此已经存在自动化技术（例如 Resy），并且不清楚 LLM 在使用该技术（无差异）之上提供什么。也就是说，您可以构建一些东西，让您可以要求 LLM 为您在 Resy 上进行预订。

—

BO*NET 基本技能定义

基本技能

促进学习或更快速地获取知识的发展能力。

内容

背景结构需要在各种不同的领域工作并获得更具体的技能。

- **阅读理解**— 理解与工作相关的文件中的书面句子和段落。
- **积极倾听**— 充分注意其他人在说什么，花时间理解所提出的要点，适当提问，不要在不适当的时候打断别人。
- **写作**— 根据听众的需要，以书面形式进行有效沟通。
- **请讲**— 与他人交谈以有效地传达信息。
- **数学**— 使用数学来解决问题。
- **科学**— 用科学的规律和方法解决问题。

过程

有助于更快速地获取各个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、结论或问题方法的优缺点。
- **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**— 选择和使用适合学习或教授新事物时的情况的培训/教学方法和程序。
- **监控**— 监控/评估您自己、其他个人或组织的绩效，以做出改进或采取纠正措施。

跨职能技能

注意：由于我们事先了解模型的编码能力，因此我们从跨职能技能列表中仅选择了编程。

- **编程**-为各种目的编写计算机程序。

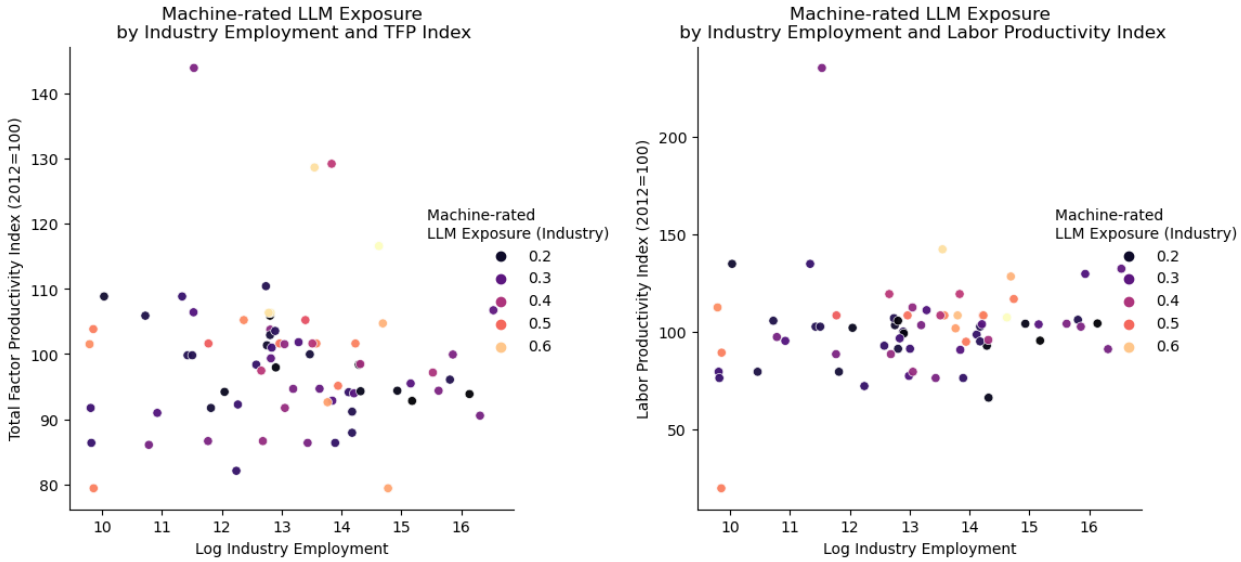
C教育

	收入中位数	Emp (000s)	H	米	H	米	H*Z*	米*Z*
无正式教育证书 高中文凭或同等学历 高等教育非学位证书 一些大学，无学位	31,900 美元	36,187	0.05	0.06	0.10	0.10	0.15	0.15
	45,470 美元	67,033	0.09	0.13	0.20	0.25	0.31	0.37
	48,315 美元	9,636	0.07	0.15	0.19	0.28	0.31	0.41
	40,970 美元	2,898	0.23	0.34	0.39	0.53	0.55	0.72
副学士学位	60,360 美元	3,537	0.12	0.14	0.31	0.36	0.49	0.59
学士学位	78,375 美元	71,698	0.23	0.17	0.47	0.51	0.70	0.84
硕士	79,605 美元	3,216	0.26	0.14	0.46	0.44	0.66	0.74
博士或专业学位	82,420 美元	5,290	0.21	0.13	0.41	0.43	0.60	0.74

表 10：职业的平均接触分数，按进入该职业所需的典型教育分组。除了曝光分数，我们还显示了每个职业的年收入中位数，以及每个组中的工人总数，以千为单位。

D 工业和生产力曝光

图 6 和图 7 分别显示了根据人类评分员和 GPT-4（基于我们的曝光量表）的 3 位数 NAICS 行业的总体就业加权相对曝光度。影响潜力存在于几乎所有行业，具有广泛的异质性。两种方法在相对暴露方面大体一致：数据处理、信息处理和医院都具有高暴露。



最近的生产率增长（全要素和劳动力）似乎也与风险敞口无关。图 D 和 D 显示自 2012 年以来的生产率增长与模型评估的当前 LLM 接触之间几乎没有关系。已经快速增长的生产性行业与风险敞口之间的高度相关性可能意味着 Baumol 成本病的恶化。换句话说，如果 LLM 可能会在不同行业中以不同方式提高生产率，那么一个问题是，最有生产力的人会变得更有生产力。由于对这些行业生产的需求缺乏弹性，生产率最高的部门占经济投入的比例将下降。我们几乎看不到任何迹象表明情况会如此。自 2012 年以来的生产力增长与 LLM 技术的接触似乎无关。

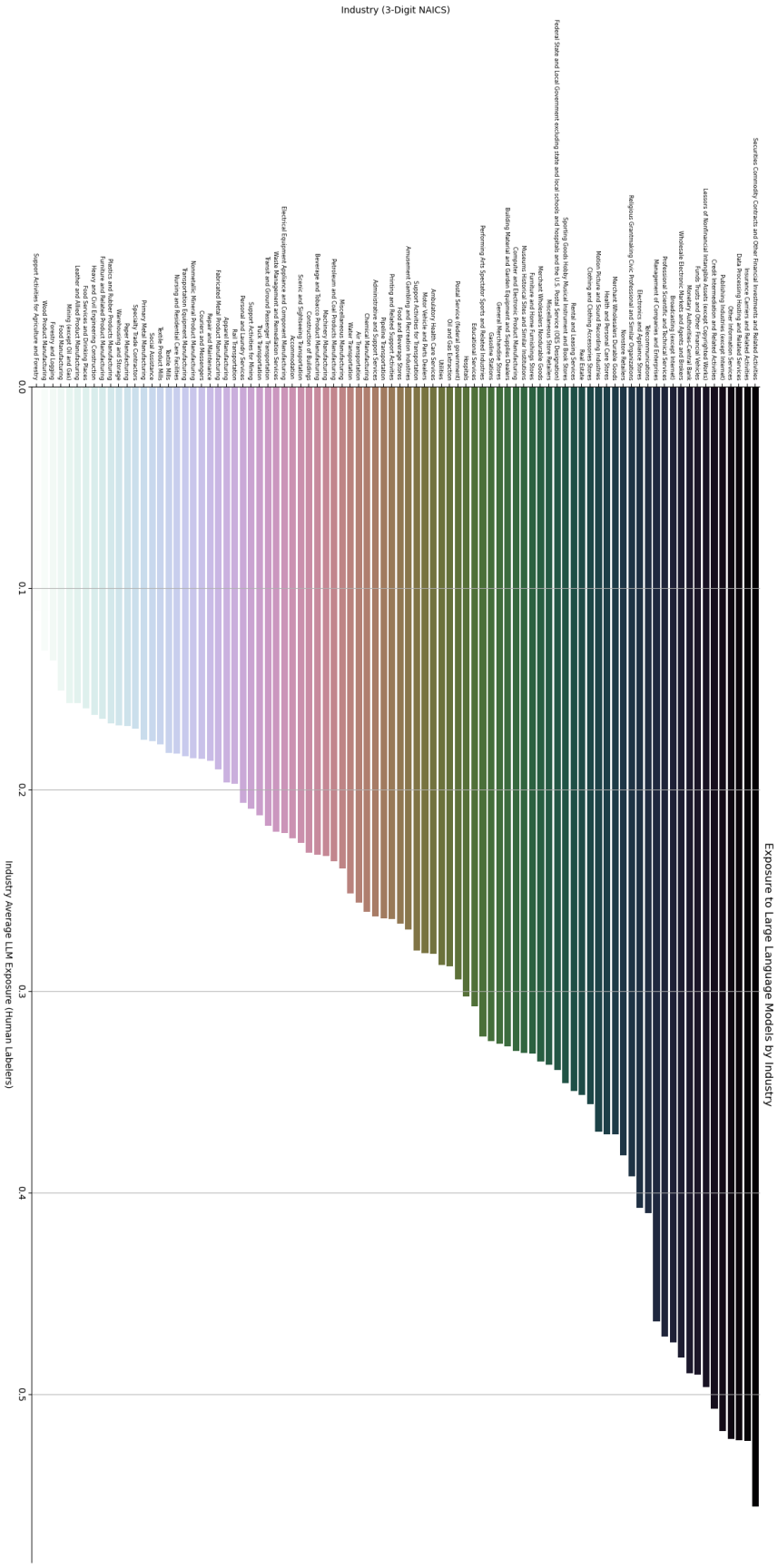

Exposure to Large Language Models by Industry
Industry Average LLM Exposure (Human Labelers)
0.0
0.1
0.2
0.3
0.4
0.5
Industry (3-Digit NAICS)
Securities Commodity Contracts and Other Financial Investments and Related Activities
Insurance Carriers and Related Activities
Data Processing Hosting and Related Services
Publishing Industries (except Internet)
Other Information Services
Couriers and Express Delivery Services
Lessors of Nonfinancial Intangible Assets (except Copyrighted Works)
Funds Trusts and Other Financial Vehicles
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
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)
Merchant Wholesalers Nondurable Goods
Furniture and Home Furnishings Stores
Miscellaneous Manufacturing
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)
Other Services (except Public Administration)
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
Administrative and Support Services
Chemical Manufacturing
Air Transportation
Water Transportation
Miscellaneous Manufacturing
Petroleum and Coal Products Manufacturing
Machinery Manufacturing
Beverage and Tobacco Product Manufacturing
Rail Transportation
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
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
Wood Product Manufacturing
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)
Oil and Gas Extraction
Food Manufacturing
Forestry and Logging
Wood Product Manufacturing
Support Activities for Agriculture and Forestry

图 6

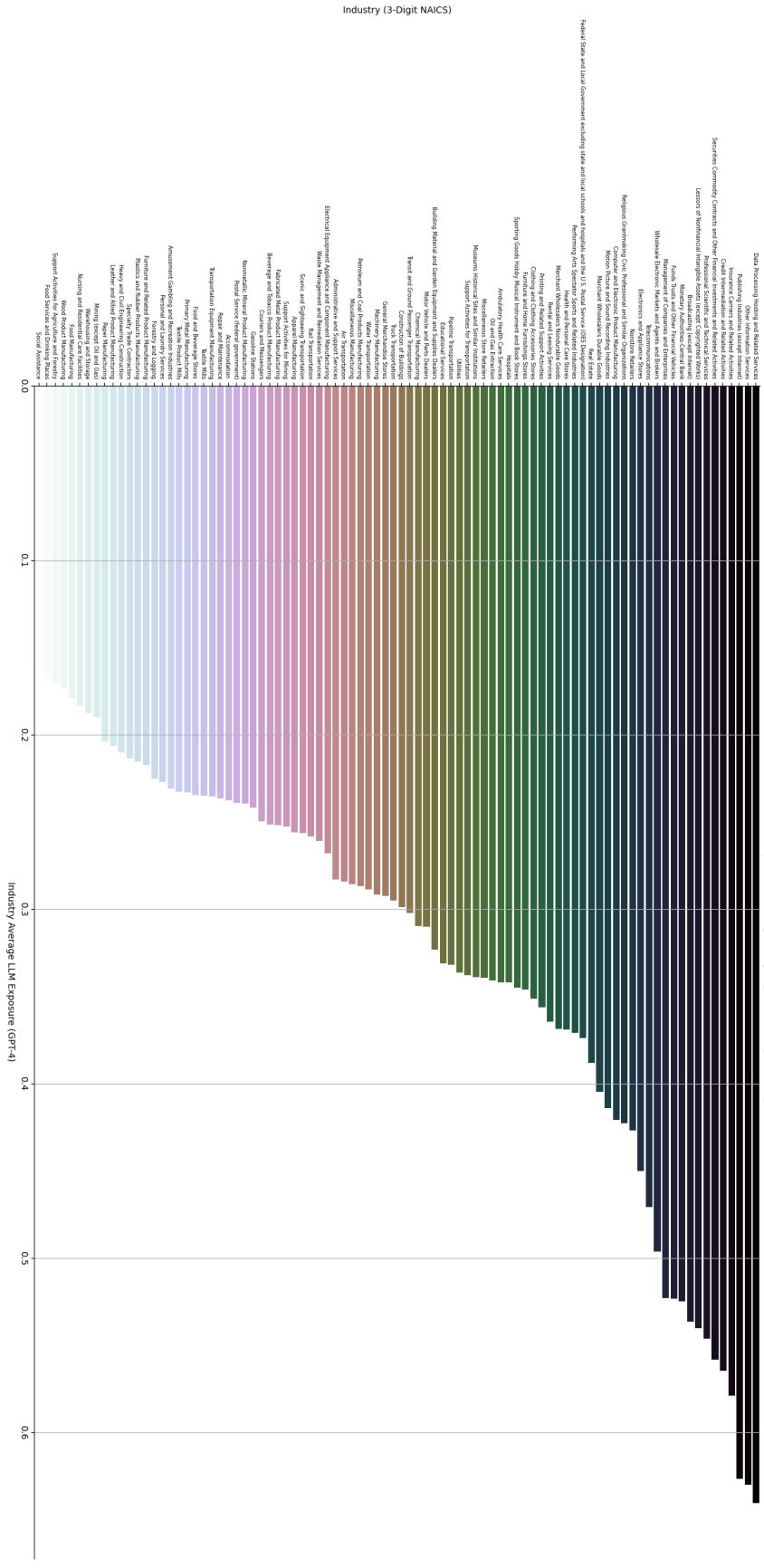

Exposure to Large Language Models by Industry
Industry (3-Digit NAICS)
Industry Average LLM Exposure (GPT-4)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Data Processing, Hosting and Related Services
Other Information Services
Publishing Industries (except Internet)
Insurance Carriers and Related Activities
Credit Intermediation and Related Activities
Securities, Commodity Contracts and Other Financial Investments and Related Activities
Professional, Scientific, and Technical Services
Lessors of Nonfinancial Intangible Assets (except Copyrighted Works)
Monetary Authorities-Central Bank
Funds, Trusts, and Other Financial Vehicles
Management of Companies and Enterprises
Wholesale Electronic Markets and Agents and Brokers
Telecommunications
Electronics and Appliance Stores
Religious, Grantmaking, Civic, Professional, and Similar Organizations
Nonstore Retailers
Motion Picture and Sound Recording Industries
Merchant Wholesalers, Durable Goods
Federal, State, and Local Government excluding state and local schools and hospitals and the U.S. Postal Service (OES Designation)
Real Estate
Performing Arts, Spectator Sports, and Related Industries
Merchant Wholesalers, Nondurable Goods
Health and Personal Care Stores
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
Construction of Buildings
Transit and Ground Passenger Transportation
Truck Transportation
General Merchandise Stores
Water Transportation
Machinery Manufacturing
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
Postal Service (federal government)
Rental and Leasing Services
Accommodation
Transportation Equipment Manufacturing
Textile Mills
Food and Beverage Stores
Textile Product Mills
Primary Metal Manufacturing
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
Paper Manufacturing
Mining (except Oil and Gas)
Printing and Related Support Activities
Warehousing and Storage
Nursing and Residential Care Facilities
Paper Manufacturing
Wood Product Manufacturing
Support Activities for Agriculture and Forestry
Food Services and Drinking Places
Social Assistance

图 7

没有任何公开任务的职业

没有标签暴露任务的职业
农业设备操作员 运动员和运动选手 汽车 玻璃安装工和维修工
公共汽车和卡车机械师和柴油发动机专家水泥泥瓦匠 和混凝土修整工 厨师，短期订单 切割机和修整机，手井架操作 员，石油和天然气 餐厅和自助餐厅服务员和调酒师助手洗碗机
疏浚操作员 电力线安装和维修人员 挖掘和装载机和吊斗铲操作员，露天采矿地板层，地毯、木材和硬瓷砖 除外 铸造模具和制芯机 助手——砖匠、砌块匠、石匠、瓷砖和大理石铺设工 助手——木匠
助手——油漆工、纸架工、泥水匠和灰泥泥瓦匠 助手——铺 管工、水管工、管道工和蒸汽装配工 助手——屋顶工
肉类、家禽和鱼类切割机和修整机摩托车技 工 铺路、铺面和夯实设备操作员打桩机操作员
倒酒器和脚轮，金属的 铁轨铺设和维护设备操作员耐火材料修理工，砖瓦工屋 顶螺栓工除外，采矿
Roastabouts、石油和天然气 屠宰者和肉类包装工 Stonemasons 锥度 轮胎修理和更换工 井口泵

表 11：我们的措施均未将任何任务标记为暴露的所有 34 种职业。

参考

Abid, A.、Farooqi, M. 和 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、伦理和社会会议论文集 , AIES '21，第 298-306 页，美国纽约州纽约市。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. 和 Autor, D. (2011a)。技能、任务和技术：对就业和就业的影响
收益。在劳动经济学手册 ，第 4 卷，第 1043–1171 页。爱思唯尔。

Acemoglu, D. 和 Autor, D. (2011b)。技能、任务和技术：对就业和就业的影响
收益。在 Ashenfelter, O. 和 Card, D. 的编辑中，劳动经济学手册 , 第 4 卷劳动经济学手册 , 第 12
章，第 1043-1171 页。爱思唯尔。

Acemoglu, D.、Autor, D.、Hazell, J. 和 Restrepo, P.（2020 年）。人工智能与工作：来自在线职位空缺的证据。
技术报告，国家经济研究局。

Acemoglu, D. 和 Restrepo, P. (2018)。人机竞赛：技术对人类的影响
增长、要素份额和就业。美国经济评论 , 108(6):1488–1542。

Acemoglu, D. 和 Restrepo, P. (2019)。自动化和新任务：技术如何取代和恢复原状
劳动。经济展望杂志 , 33(2):3–30。

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

Acemoglu, D. 和 Restrepo, P. (2022b)。任务、自动化和美国工资不平等的加剧。计量经济学 ，
90(5):1973–2016。

Aghion, P.、Jones, BF 和 Jones, CI (2018)。人工智能与经济增长。在这
人工智能经济学：议程 , 第 237–282 页。芝加哥大学出版社。

Agrawal, AK、Gans, JS 和 Goldfarb, A.（2021 年）。人工智能的采用和全系统的变革。技术报告，
国家经济研究局。

Arntz, M.、Gregory, T. 和 Zierahn, U.（2017 年）。重新审视自动化的风险。经济学快报 ，
159:157–160。

Autor, D.、Chin, C.、Salomons, AM 和 Seegmiller, B. (2022a)。新领域：起源和内容
新作品，1940-2018。技术报告，国家经济研究局。

Autor, D.、Mindell, DA 和 Reynolds, EB (2022b)。未来的工作：创造更好的工作
智能机器时代 . 麻省理工学院出版社。

Autor, DH、Katz, LF 和 Kearney, MS (2006)。美国劳动力市场的两极分化。美国人
经济评论 , 96(2):189–194。

Autor, DH、Levy, F. 和 Murnane, RJ (2003)。近期技术变革的技能内容：
实证探索。经济学季刊 , 118(4):1279–1333。

Babina, T.、Fedyk, A.、He, A. 和 Hodson, J.（2021 年）。人工智能、企业成长和产品
创新。FirmGrowth 和产品创新（2021 年 11 月 9 日） .

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., 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. 和 Kaplan, J. (2022)。通过从人类反馈中强化学习来训练一个有用且无害的助手。arXiv:2204.05862 [cs]。

Baumol, WJ (2012)。成本病：为什么计算机变得更便宜而医疗保健却没有 . 耶鲁大学按。

Benzell, SG、Kotlikoff, LJ、LaGarda, G. 和 Ye, VY (2021)。模拟内生全局自动化。工作论文 29220，国家经济研究局。

J. 贝森 (2018)。人工智能和工作：需求的作用。在人工经济学情报：议程 , 第 291–307 页。芝加哥大学出版社。

美国劳工统计局 (2022)。按详细职业划分的就业。

美国劳工统计局 (2023a)。人口统计学特征（cps）。

美国劳工统计局 (2023b)。职业展望手册 az 指数。

Bommasani, R., Hudson, DA, Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, MS, Bohg, J., Bosselut, A., Brunskill, E., 等人。(2021)。关于基础模型的机会和风险。arXiv 预印本 arXiv:2108.07258 .

T. 布雷斯纳汉 (2019)。人工智能技术和聚合增长前景。

Bresnahan, T.、Greenstein, S.、Brownstone, D. 和 Flamm, K.（1996 年）。技术进步和共同发明在计算和计算机的使用中。布鲁金斯经济活动论文。微观经济学 , 1996:1-83。

Bresnahan, TF (1999)。计算机化和工资分散：分析性的重新解释。经济的杂志 , 109(456):390–415。

Bresnahan,TF、Brynjolfsson,E. 和 Hitt,LM(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, TF 和 Trajtenberg, M. (1995)。通用技术“增长引擎”？ 杂志计量经济学 , 65(1):83–108。

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, JD, Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G.、Askell、A. 等人。（2020）。语言模型是少样本学习者。神经信息处理系统的进展 , 33:1877–1901。

Brynjolfsson, E.、Frank, MR、Mitchell, T.、Rahwan, I. 和 Rock, D. (2023)。量化分布机器学习对工作的影响。即将到来 .

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

Brynjolfsson, E.、Mitchell, T. 和 Rock, D.（2018 年）。机器可以学习什么，这对什么意味着什么职业和经济？AEA 论文和会议记录 , 108:43–47。

Brynjolfsson, E.、Rock, D. 和 Syverson, C.（2021 年）。生产力 j 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。

蔡斯 H. (2022)。郎链。

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, HP d. O., Kaplan, J., Edwards, H., 布尔达, Y., 约瑟夫, N.，布罗克曼，G.，等。(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. 和 Tambe, P. (2022)。Innovae：用于理解专利和创新的生成式人工智能。可在 SSRN .

阿肯色州周 (2023)。为什么 ChatGPT 是有史以来增长最快的网络平台 | 时间。

Cockburn, IM、Henderson, R. 和 Stern, S.（2018 年）。人工智能对创新的影响：一个探索性分析。在人工智能经济学：议程 , 第 115-146 页。芝加哥大学出版社。

J. 康斯坦茨 (2023)。据称，近三分之一的白领工人尝试过 chatgpt 或其他人工智能程序一项新的调查。

宾夕法尼亚州大卫 (1990)。发电机和计算机：现代生产力的历史视角悖论。美国经济评论 , 80(2):355–361。

Devlin, J.、Chang, M.-W.、Lee, K. 和 Toutanova, K.（2019 年）。Bert：深度双向预训练语言理解的转换器。ArXiv , abs/1810.04805。

Dixon, J.、Hong, B. 和 Wu, L.（2021 年）。机器人革命：对管理和就业的影响公司。管理科学 , 67(9):5586–5605。

Feigenbaum, JJ 和 Gross, DP (2021)。组织摩擦和自动化回报递增：二十世纪 at&t 的教训。技术报告，国家经济研究局。

Felten, E.、Raj, M. 和 Seamans, R. (2023)。像 chatgpt 这样的语言建模器将如何影响职业和行业？arXiv 预印本 arXiv:2303.01157 .

Felten, EW、Raj, M. 和 Seamans, R.（2018 年）。一种将人工智能的进步与职业能力。AEA 论文和会议记录 , 108:54–57。

CB 弗雷 (2019)。技术陷阱。在技术陷阱 . 普林斯顿大学出版社。

Frey,CBandOsborne,MA(2017)。就业的未来：工作对计算机化有多敏感？技术预测和社会变革 , 114(C):254–280。

Goldfarb, A.、Taska, B. 和 Teodoridis, F. (2023)。机器学习可以成为通用技术吗？A使用来自在线职位发布的数据比较新兴技术。研究政策 , 52(1):104653。

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

Grace, K.、Salvatier, J.、Dafoe, A.、Zhang, B. 和 Evans, O.（2018 年）。人工智能什么时候会超过人类的表现？来自人工智能专家的证据。人工智能研究杂志 , 62:729–754。

Hernandez, D.、Kaplan, J.、Henighan, T. 和 McCandlish, S.（2021 年）。传输的缩放法则。arXiv 预印本 arXiv:2102.01293 .

JJ 霍顿 (2023)。作为模拟经济主体的大型语言模型：我们可以从 homo 中学到什么硅胶？arXiv 预印本 arXiv:2301.07543 .

黄，M.-H。和 Rust，RT (2018)。人工智能服务。服务研究杂志 , 21(2):155–172。

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

Katz, LF 和 Murphy, KM (1992)。相对工资的变化，1963-1987：供给和需求因素。
经济学季刊 , 107(1):35–78。

Khlaaf, H.、Mishkin, P.、Achiam, J.、Krueger, G. 和 Brundage, M. (2022)。危害分析框架
代码合成大型语言模型。

Klinova, K. 和 Korinek, A. (2021)。艾与共享繁荣。在AIES 2021 - 2021 年论文集
AAAI/ACM 人工智能、伦理和社会会议 .

Kogan, L.、Papanikolaou, D.、Schmidt, LDW 和 Seegmiller, B.（2021 年）。技术，特定年份
人力资本和劳动力转移：来自专利与职业联系的证据。工作论文 29552，国家经济研究局。

Korinek, A. (2023)。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. 和 Stiglitz, JE (2018)。人工智能及其对收入分配的影响
和失业。在人工智能经济学：议程 , 第 349–390 页。芝加哥大学出版社。

Lipsey,RG,Carlaw,KI 和 Bekar,CT (2005)。经济转型：通用技术
和长期经济增长 . 牛津。

Meindl, B.、Frank, MR 和 Mendonça, J.（2021 年）。职业接触第四代技术
工业革命。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. 等人。(2023)。增强语言模型：一项调查。arXiv 预印本
arXiv:2302.07842 .

Moll, B.、Rachel, L. 和 Restrepo, P. (2021)。不平衡增长：自动化对收入和财富的影响
不等式。SSRN 电子期刊 .

Mollick, ER 和 Mollick, L. (2022)。人工智能聊天机器人启用的新学习模式：三种方法和
作业。可在 SSRN .

Noy, S. 和 Zhang, W. (2023)。人工生成生产力影响的实验证据
智力。可在 SSRN 4375283 获得 .

O*NET (2023)。O*net 27.2 数据库。

开放人工智能 (2022)。介绍聊天。

开放人工智能 (2023a)。Gpt-4 系统卡。技术报告，OpenAI。

开放人工智能 (2023b)。Gpt-4 技术报告。技术报告，OpenAI。

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

Peng, S.、Kalliamvakou, E.、Cihon, P. 和 Demirer, M. (2023)。人工智能对开发人员生产力的影响：
来自 github copilot 的证据。arXiv 预印本 arXiv:2302.06590 .

Radford, A.、Wu, J.、Child, R.、Luan, D.、Amodei, D.、Sutskever, I. 等。（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)。四分之一的公司已经用 chatgpt 取代了工人。

D. 洛克 (2019)。工程价值：技术人才和人工投资的回报智力。可在 SSRN 3427412 获得 .

Schick, T.、Dwivedi-Yu, J.、Dessì, R.、Raileanu, R.、Lomeli, M.、Zettlemoyer, L.、Cancedda, N. 和 Scialom, T. (2023)。Toolformer：语言模型可以自学使用工具。arXiv 预印本 arXiv:2302.04761 .

Schramowski, P.、Turan, C.、Andersen, N.、Rothkopf, CA 和 Kersting, K. (2022)。大型预训练语言模型包含类似人类的判断正确和错误行为的偏见。自然机器智能 , 4(3):258–268。

Shahaf,D.andHorvitz,E.（2010）。人机计算的广义任务市场。诉讼程序 AAAI 人工智能会议 .

Singla, AK、Horvitz, E.、Kohli, P. 和 Krause, A.（2015 年）。学习雇佣团队。在AAAI 会议人类计算与众包 .

Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Kim, JW、Kreps, S.、McCain, M.、Newhouse, A.、Blazakis, J.、McGuffie, K. 和 Wang, J.（2019 年）。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. 和 Wingate，D. (2022)。一种信息论方法，用于在没有地面真值标签的情况下提示工程。在 计算语言学会第 60 届年会论文集（第 1 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., 等人。(2022)。Lamda：对话应用程序的语言模型。arXiv 预印本 arXiv:2201.08239 .

Tolan, S.、Pesole, A.、Martínez-Plumed, F.、Fernández-Macías, E.、Hernández-Orallo, J. 和 Gómez, E. (2021)。衡量人工智能的职业影响：任务、认知能力和人工智能基准。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。

韦伯，M。（2020）。人工智能对劳动力市场的影响。工作论文，斯坦福大学。

Weidinger, L. 等人。(2021)。语言模型危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, LA、Rimell, L.、Isaac, W.、Haas , J.、Legasick, S.、Irving, G. 和 Gabriel, I.（2022 年）。语言模型带来的风险分类。在2022 年 ACM 公平性、问责制和透明度会议 , FAccT '22，第 214–229 页，美国纽约州纽约市。计算机协会。

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