arXiv:2603.02076v1 [econ.GN] 2 Mar 2026

当人工智能评判你的工作：算法评估的隐性成本*

David Almog†、Lucas Lippman‡和 Daniel Martin§

2026年1月5日

摘要

我们通过一项包含真实工作任务的在线实验，研究劳动者在知道自己的工作将由人工智能而非人类评估时，是否会改变行为。我们发现，被指定由人工智能评估者评估的个体产出数量更多。然而，在控制产出数量后，无论采用人类评估还是大语言模型评分来衡量，产出质量都较低。我们还发现，与知道工作将由人类评估相比，知道工作将由人工智能评估时，劳动者更可能使用外部工具，包括大语言模型。不过，外部工具使用的增加似乎无法解释不同处理组之间在产出数量或质量上的差异。

*已获得加州大学圣塔芭芭拉分校机构审查委员会的批准。

†凯洛格管理学院，西北大学，david.almog@kellogg.northwestern.edu。

‡Walmart Connect。§加州大学圣塔芭芭拉分校，

danielmartin@ucsb.edu。

1 引言

生成式人工智能（GenAI）的兴起主要由大语言模型（LLMs）推动，大幅拓展了人工智能可评估的任务范围，例如书面文件和音频转录文本的语气与内容。因此，企业和机构日益开始考虑是否使用人工智能进行评估。在工作场所，人工智能已开始用于绩效评估和招聘（Goergen, Bellis, and Klesse 2025）。在学术领域，人们正考虑将其用于作业评分（D. W. Zhang, Boey, Tan, and Jia 2024；Flodén 2025）以及申请筛选[1]。在经济实验中，大语言模型正被用于激励绩效（Conlon 2025）。

为什么企业、机构和实验人员会选择人工智能评估，而非人工评估？一个关键原因是，人工智能评估的成本可能更低。当人工评估需要领域专家（例如管理者）参与时，这一点尤其明显。人工智能评估也可能更快捷，能够即时提供反馈、更快地完成分析并迅速付款。最后，人工智能评估可能更加一致，因为可以固定训练好的模型，从而确保每位劳动者或每份文档都接受相同的评估。

事实上，我们的研究表明，仅凭节省时间和成本，就足以产生实施人工智能评估的强烈激励。在我们的实验中，使用人工智能为所有受试者的产出评分，花费了 11.67 美元，另加合著者编写 Python脚本的时间；该脚本用一天时间完成编码和运行。相比之下，人工评分由三名研究生完成，平均耗时 54 小时（将近七个 8 小时工作日），总成本为 6,480 美元。

虽然人工智能评估可能更具成本效益、速度也更快，但如果转而采用人工智能评估会对劳动者产出产生负面影响，企业和机构可能会对放弃人工评估有所顾虑。[2]一个尚待解答的问题是，人工智能评估下的劳动者产出是否与人工评估下的劳动者产出存在根本差异。这个问题对于理解企业何时以及在多大程度上会采用人工智能评估，以及采用人工智能评估会带来哪些影响，至关重要。

本文通过一项在线实验研究人工智能评估对劳动者产出的影响。在该实验中，受试者通过为图像提供图像描述获得报酬。这是在线劳动者经常执行的一项任务，包括通过平台完成计件工作的劳动者，以及接受更为标准的工作合同、承担此类任务的劳动者。[3] 为图像添加说明文字是各行业普遍存在的一项任务：社交媒体经理为帖子添加说明文字，保险专员为受损商品的理赔申请添加说明文字，标注员为机器学习模型的训练数据添加说明文字，图书馆员和档案管理员为馆藏资料添加说明文字以便搜索工具检索，放射科医生则为X光片添加说明文字以辅助

[1]参见 https://apnews.com/article/ai-chatgpt-college-admissions-essays-87802788683ca4831bf1390078147a6f。[2]他们可能犹豫的其他原因包括负面宣传、法律风险或员工士气。例如，算法评估可能会被认为不够公平（Mok、Nanda 和 Anderson 2023）。[3]参见 https://www.indeed.com/q-image-captioning-jobs.html。

解读。[4] 尽管人工智能越来越多地用于图像描述，但人工智能生成的描述仍存在相当严重的质量问题（Sarhan 和 Hegelich 2023），因此人类仍会定期受雇来完成这项任务。

在我们于 Prolific 平台开展的实验中，我们要求受试者为 20 张日常场景图像撰写 150 至 400 个字符的说明文字，例如摊贩挤压橙子或孩子们拿着滑板的场景。我们告知受试者，他们的说明文字将根据三项标准进行评估（有趣、用心和准确），且整个实验期间这些标准保持不变。[5] 每当受试者的图像描述在某张图像对应的排名中位居前 30% 时，便可获得$1 奖金，因此每位受试者最多可赚取$20 ，另加$8出席费；这一金额远高于 Prolific 平台的标准时薪（Palan 和 Schitter，2018）。我们采用被试间设计，随机安排受试者由人工评估者或 ChatGPT（研究开展时领先的大语言模型）评估其说明文字。重要的是，我们告知受试者由谁评估他们，但未告知他们还存在另一种处理条件。受试者被分配的评估者（“分配至人工评估者”或“分配至 ChatGPT”）在全部 20 张图像中保持不变。如果受试者被分配给人工评估者，我们还如实告知他们，该评估者是一名大学毕业生。

我们发现，被分配给人工智能评估者后，产出数量会增加 27.8% 个标准差，这一指标以说明文字的长度衡量。然而，在控制产出数量后，我们发现，当受试者被分配由人工智能评判时，说明文字的质量会下降。为了评估说明文字的质量，我们采用了几项基准：博士生评审员给出的数值评分，以及 GPT-4o 在不同温度参数和提示词下给出的数值评分。在控制产出数量后，被分配给人工智能评估者会使评分在各项质量基准下下降 12.1% 至 20.5% 个标准差。纳入人口统计学控制变量和图像固定效应，并在个体层面聚类计算标准误后，所有基准下的这些降幅在 5% 水平均具有统计显著性。

我们还发现，与由人类评估相比，由人工智能评估其作品时，个体使用外部工具的可能性高出 93.3%（9.7 个百分点）。然而，借助外部工具完成的说明文字往往质量更高，因此，外部帮助使用率上升并不能解释人工智能评估组说明文字质量较低的现象。我们无法直接观察外部工具的使用情况，因此采用了几种方法来识别受试者是否使用了外部工具，其中主要的代理指标是粘贴文本。

受试者正在使用哪些外部工具？一种可能性是，在接受人工智能评估时，他们增加了对人工智能工具的使用。为帮助确定这一点，参与者被

[4]除了作为一项常规工作任务之外，提供书面描述也是一项标准的教育活动。[5]这些标准的选择将在实验设计部分详细说明。

要求披露他们是否使用过大语言模型来帮助撰写描述；如果其回答与人工智能检测软件的预测相符，则可获得金钱奖金。这提供了一种新颖的激励式引出方法，用于识别人工智能使用情况。我们利用这一方法表明，在接受人工智能评估时，人工智能工具的使用有所增加，随后证明，这一指标可以作为粘贴指标的有益补充。

本文其余部分安排如下。第1.1 节回顾经济学中的人工智能相关文献。第2 节介绍我们的实验设计和关键设计选择。第3 节考察评估者分配对写作习惯、质量和外部协助使用的处理效应。在第4节，我们讨论研究发现的解释及其意义。最后，第5 节作简要总结。

1.1 文献综述

本文对经济学的多个研究领域有所贡献，这些领域探讨了人工智能使用对社会的影响。

生成式人工智能的兴起势必以前所未有的规模冲击劳动力市场。Eloundou、S. Manning、Mishkin 和 Rock（2024）估计，美国 80% 的劳动者所承担的任务中，至少有 10% 可能受到大语言模型的影响；其中 19% 的劳动者所承担的任务中，超过一半可能受到影响。这一潜力已开始显现。Humlum 和 Vestergaard（2025）利用全面的丹麦登记数据，对劳动者开展了一项大规模调查，发现至少半数受访者曾在工作中使用 ChatGPT。类似地，Bick、Blandin 和 Deming（2024）基于一项针对美国 18 至 64 岁人口的代表性调查报告称，截至 2024年8月，39% 的受访者使用过生成式人工智能。[6] 迄今为止，绝大多数研究都聚焦于让劳动者使用生成式人工智能工具所带来的生产率提升（例如，Noy 和 W. Zhang2023；Brynjolfsson、Li 和 Raymond2025；Peng、Kalliamvakou、Cihon 和 Demirer2023；Choi 和 Schwarcz2025；Otis、Clarke、Delecourt、Holtz 和 Koning2025）。然而，人工智能在工作场所环境的其他方面也具有巨大的采用潜力（Ludwig、Mullainathan 和 Rambachan2024），其中一种应用便是将其用作评估机制。

Autor、Chin、Salomons 和 Seegmiller（2024）以及 Acemoglu 和 Restrepo（2019）讨论了技术冲击如何同时促成自动化和新任务的创造。在此基础上，Ide 和 Talamas（2025）对知识经济的经典模型进行了调整，以揭示人工智能可能如何重塑层级制企业的组织层面影响。为了准确评估哪些任务面临人类劳动力替代的风险，以及企业的这种重组将如何展开，有必要评估将人工智能整合进工作流程所产生的一般均衡效应，包括人类对其存在的反应。我们的论文通过考察 AI 采用背后的激励因素，为这一文-

[6]此后，这些数字可能有所增加。

章作出了贡献。我们引入了一种新的权衡，用于指导何时将人工智能用作评估工具的决策，并证明人类劳动者对于受到不同主体（人工智能与人类）监控的反应，可能会显著影响这一决策。

我们也为有关监控的文献作出了贡献，尤其是探讨人们对受到监控所产生的行为反应的研究。与委托—代理理论的预测相反，Frey（1993）提出，监控可能会降低内在动机，从而对劳动者的努力产生负面影响。Dickinson 和 Villeval（2008）随后对这一“挤出”现象进行了细致检验，并提供证据表明，监控强度超过某一阈值后，努力程度会下降。Nagin、Reb- itzer、Sanders 和 Taylor（2002）基于呼叫中心的现场实验发现，劳动者可能会对监控作出策略性反应，这与“理性作弊者”模型相符。如果劳动者会因监控而调整行为，而且人与计算机玩家进行策略互动时的表现不同于与其他人互动时的表现[7]，那么，预期人们对人工智能评估和人工评估作出不同反应是合理的。本文为这一差异提供了证据，并探讨了这些差异可能如何产生。

与我们的研究最密切相关的两篇论文是 Almog、Gauriot、Page 和 Martin（2024）以及 Corgnet、Hernán-González 和 Mateo（2023）。前一篇论文提供了证据，表明引入人工智能评估影响了职业网球裁判所犯错误的程度和性质。然而，在引入人工智能评估之前，并没有对裁判的判罚进行人工评估，因此该研究情境无法用于回答我们关注的问题。为了真正探讨劳动者在人工判断和人工智能判断之间的反应是否存在差异，我们实施了一项任务，使两种监控安排都能够以可比的方式进行。后一篇论文研究了用算法取代人类对同伴效应的影响。他们的研究结果表明，引入算法会通过减轻社会压力，显著影响劳动者的表现；在他们的研究情境中，社会压力是影响表现的重要因素。一个关键区别在于，在他们的研究中，算法充当同事的替代者，并通过表现不佳所带来的尴尬或自私行为的展现来影响社会压力。而在我们的论文中，人工智能的实施方式有很大不同：它承担质量控制主管的角色，对劳动者进行监控。尽管尴尬在我们的研究情境中也可能发挥作用（正如在人际互动的大多数情境中很可能如此），但我们的重点不再是应对自私行为，而是探讨其他问题，例如劳动者何时更有可能寻求外部帮助。总的来说，这两篇论文与我们的研究都秉持这样的观点：有效的自动化必须在评估中考虑行为因素。[8]

[7]March（2021）综述了 162 项实验研究，为此提供了证据。[8]Camerer（2019）强调，将行为经济学的见解融入企业和机构中的人工智能—人类互动，对于改善这类互动至关重要。

最后，我们通过引入一种能够诱导人工智能使用的激励方法，并将其与其他几项收集到的外部协助指标一同进行检验，为实验经济学工具箱作出了贡献。生成式人工智能为改进实验实践的许多领域提供了机会（参见 Korinek 2023；Charness、Jabarian 和 List2023；S. Zhang、Xu 和 Alvero 2025），但与此同时也带来了新的挑战，因为参与者可能获得帮助，甚至将实验任务委托给他人。Veselovsky、Ribeiro、Cozzolino、Gordon、Rothschild 和 West（2025）利用 Prolific 上的一项文本摘要任务估计，大约 30% 的参与者使用了大语言模型。他们的研究结果表明，限制复制粘贴可以将大语言模型的使用量减少一半。结合我们诱导获得的 ChatGPT 使用指标与文本粘贴检测器，我们表明，人工评估也是抑制大语言模型使用的有效方式。我们的研究结果与先前研究一致，即人类的存在能够有效遏制不诚实行为（Cohn、Gesche 和 Maréchal 2022）。

2 实验

2.1 实施

我们于[9] 2024年11月开展了预注册的在线实验。实验网页使用基于 Python 的 Web 框架 Flask 构建，并托管于*PythonAnywhere* 平台。实验使用了 OpenAI 的应用程序接口，并遵循了大语言模型研究领域的许多当前最佳实践（Filippas、Horton 和 B. S. Manning 2025）。在应用程序接口集成方面，我们使用了 OpenAI 官方的 Python客户端库。每次应用程序接口调用都是独立请求，确保评估之间不保留对话历史或上下文，并且每个请求都包含正确的系统提示词、图像和图像描述。我们招募了 208 名参与者[10]，他们来自 Prolific 的参与者群体；我们仅选取居住在美国的参与者，以确保他们具备一定的英语熟练度，能够完成写作任务。

同意参与实验后，参与者立即收到了总体说明。他们得知需要完成 20 轮，每轮都要用 150 至 400 个字符描述一幅图像。研究人员提供了四幅示例图像，并说明描述将依据三项标准进行评估：有趣、用心和准确。参与者还被告知，对于每幅图像，只要其描述在其他 Prolific 参与者提交的描述中排名前 30%，便可获得 1 美元的“成功”奖金。随后，他们得知“成功”奖金将由 [ChatGPT/一名研究生], 来决定，但每位参与者只会被告知其中一位评估者，从而维持

[9]预注册计划可见于 https://aspredicted.org/qc3fn5.pdf[10]103 名和 105 名参与者分别被分配至人工评估组和ChatGPT处理组。

实验的被试间设计。最后，参与者完成了一轮练习，他们知道这一轮不会计入报酬，但可以帮助他们熟悉任务和界面。实验说明和界面的截图见附录 B.1。

在阅读说明并完成练习轮次后，参与者的任务是为描绘日常生活场景的20幅图像撰写图像描述。[11] 界面包含字符计数器，并在底部持续显示提醒，说明他们的评估者是谁，以及他们接受评估所依据的标准。完成20轮图像描述撰写后，参与者受邀参加一项关于其任务体验的可选调查问卷。最重要的是，研究人员询问他们是否使用过类似ChatGPT的工具来协助完成任务。起初，这一问题不附带激励。随后，我们实施了激励式方法：如果参与者的回答与人工智能检测软件的预测一致，他们将额外获得1美元。

参与者获得了8美元的出席报酬，并有机会根据表现额外赚取最高20美元。图像描述由分配给他们的评估者评分；如果某条图像描述在针对特定图像的排名中位列前30%，参与者将获得1美元奖金。[12] 实验平均持续75分钟；根据设计，平均报酬为14美元，显著高于Prolific针对这一时长研究所建议的报酬率。

2.2 设计选择

在本节中，我们 讨论我们在设计实验时面临的几个主要选择 .

选择 *3* 项评估标准：有趣、用心和准确。 我们认为，选择多项标准有助于防止参与者试图钻任务的空子，并确保图像描述既符合标准，又能有效描述图像。在选择一组三项相辅相成的标准时，我们决定利用大语言模型的一些新能力，采用主观标准（有趣 和 用心），同时保留客观测量指标，例如准确性。在试点中，我们随机安排受试者看到三项客观标准或三项主观标准。分配不同评估者所产生的处理效应相当相似，[13] 因此我们将客观标准和主观标准合并为一组，并仅关注评估者差异。

使用 *ChatGPT* 模型 *GPT-4o*。在我们开展实验时，它是 ChatGPT 最先进的模型。此外，ChatGPT 是最受欢迎的大语言模型，这增加了

[11]这 20 张图像收录于附录 B.3。[12]由于我们的质量指标采用离散数值，处于门槛值的多位参与者获得了相同的评分。为确保公平性，我们对门槛值处的奖金进行随机分配，使 30% 的参与者获得奖金。[13]请参阅 C.1。

人们熟悉它的可能性。我们使用的模型具备解读图像的能力，这一功能在实验进行前不久推出，对于此项任务的人工智能评估至关重要。

提示词。 在测试过程中，我们要求 ChatGPT 详细说明其评分依据。我们发现，当提示 ChatGPT 同时依据三项标准为给定的图像描述评分时，它有时会走捷径，对多项标准给出相同的评分，并提供相同的理由。为解决这一问题，我们针对每条图像描述分别设置了三个提示词，每个提示词对应一项具体标准。此外，我们注意到，温度设置较高时，ChatGPT 偶尔会偏离评分指南。为减少这种情况，我们将温度设为 0.75。如提示词所示，我们提供了系统而详细的说明，明确列出了任务、ChatGPT 的角色、评分标准、步骤和回复格式。提供给 ChatGPT 和人工评估者的完整提示词见附录 B.2。

奖金激励。 我们决定根据同一评估者内部的相对表现发放奖金。每当某条图像描述在某张图像的评选中排名进入前 30% 时，我们便发放奖金。这样就能排除与人们对评估者宽松度的看法有关的解释，我们将在第 4 节中对此作更详细的讨论。相对表现报酬可能引发一种担忧：参与者会试图推断其他人的表现。不过，我们认为，消除某些机制渠道所带来的益处远远超过这一担忧；而且我们没有证据表明这一问题在本次实验中发挥了重要作用。另一种设计是在图像描述达到满意度阈值时发放奖金（例如，得分达到 10 分制中的 7 分或以上）。试点阶段采用了这种方法，我们的结果在该设计下似乎同样稳健。

未征询参与者对偏好评估者的支付意愿。 在试点中，我们征询了参与者为其偏好的评估者支付的意愿（支付意愿，WTP），并在一定程度上根据这些偏好，将参与者分配给人工评估者或 ChatGPT评估者。即使按照参与者偏好的评估者对其进行区分，我们的结果，尤其是关于工作质量的结果，仍然稳健（C.2）。然而，工作场所环境通常会让人们接受指定的评估方法，而不论其个人偏好，因此我们决定转向包含这一特点的设计。不过，探究人们对偏好评估者的偏好，是未来研究中一个有趣的问题。

文本粘贴管理。 我们选择不限制粘贴，因为这样做可能适得其反，无意中促使参与者尝试通过钻任务空子或采取其他方式来应对任务。不过，我们认为这也是一个有趣的未来研究方向。

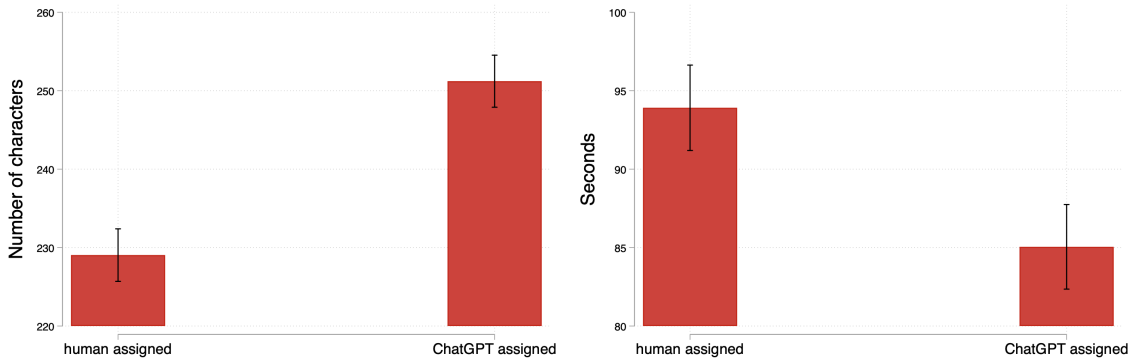
3 结果

3.1 写作习惯会随评估者而变化吗?

我们首先通过分析预注册测量的写作习惯，考察劳动者被分配给不同评估者时其行为是否发生变化：回复长度、回复时间、粘贴和擦除。这些都是可以客观量化的变量（代表真实值的真实指标）。尽管包含粘贴文本的回复在这些指标上以多种方式不同于不含粘贴文本的回复，但我们将把这一比较留到后文讨论。在此，我们聚焦于总体结果。

3.1.1 回复长度和时间

研究写作产出的一个自然起点，是考察回复长度和写作所用时间。我们发现，人工评估者评估的参与者平均会写出更短的说明文字，但耗时更长。由人工评估者评估时，平均回复长度为229 个字符；由 ChatGPT 评估时则为251 个字符（均值双侧检验的 p 值为 < 0.0001）。至于回复时间，分配给人工评估者时，平均用时为94 秒；分配给 ChatGPT评估者时则为 85 秒（均值双侧检验的 p 值为 < 0.0001）。[14] 图1以可视化方式呈现了这些发现，并凸显了结果具有很强的统计显著性。



（a）回复长度。　　（b）回复时间。

图1：按处理方式划分的平均图像描述回复长度和时间。

[14]由于存在极端离群值，我们对时间变量按第5和第95百分位数进行了缩尾处理。不进行此调整时，结果仍然一致

3.1.2 粘贴与擦除

借助为本实验定制的网页功能，我们收集了两个有助于了解写作过程的变量：粘贴文本和删除字符数。我们创建了一个虚拟变量：如果参与者在撰写说明文字的任何阶段粘贴过字符，该变量就取值为 1。我们将更详细地研究参与者粘贴文本时会发生什么，但如前所述，我们将粘贴作为使用外部帮助的主要代理指标。我们发现，不同处理条件下的粘贴习惯差异显著：由 ChatGPT 评估的参与者中，有 20.1% 使用了粘贴；而由人工评估者评估的参与者中，这一比例为 10.4%（双侧 t 检验的 p 值为 < 0.0001）。另一方面，如图A.1所示，处理对擦除没有影响。

3.2 质量差异?

我们记录了四项写作习惯变量中三项因评估者而异的显著处理效应。鉴于图像描述的输出以及撰写过程已被证明存在差异，因此图像描述的质量是否也存在差异，便值得探究。与此前各变量相比，衡量质量的关键区别在于缺乏真实值，即便是对于<si id='1'>准确性</si>这一被认为最客观的评估标准也是如此。虽然缺乏真实值可能会被视为一项挑战，但这反而增加了现实性，因为评估文本与许多其他工作相关任务一样，很少涉及明确的真实值概念。

我们请 ChatGPT 和人工评估者为每条图像描述评分，无论其所属的处理条件如何。两类评估者收到相同的说明，并分别对每条图像描述的各项标准进行评分，以尽量减少溢出效应。每项标准的得分为 0 至 3 分，一张图像的总得分最高为 9 分。尽管报酬奖金根据指定评估者给出的评分发放，我们仍收集了两类评估者对所有图像描述的评价。这种做法产生了两项质量指标：一项来自 ChatGPT，另一项来自人工评分者。

我们首先分析ChatGPT给出的得分。我们发现，人工评估者处理组的参与者（即被分配人工评估者的参与者）从ChatGPT获得的平均得分高于ChatGPT评估者处理组的参与者（即被分配ChatGPT评估者的参与者）。[15] 图2显示，人工评估者处理组的平均得分为6.32，而ChatGPT评估者的平均得分为6.05（双侧 t 检验的p 值为0.0001）。根据ChatGPT，在认为将由人工评估自己的印象下工作的参与者，最终写出的说明文字质量高于那些认为自己将

[15]这一结果对于温度参数的变化，以及使用全部三项标准进行综合评分还是分别评分，均具有稳健性。

由ChatGPT评估。这一结果足够有力，足以克服ChatGPT处理组参与者可能具备的潜在优势，即他们可以针对实际评估者来调整自己的说明文字。

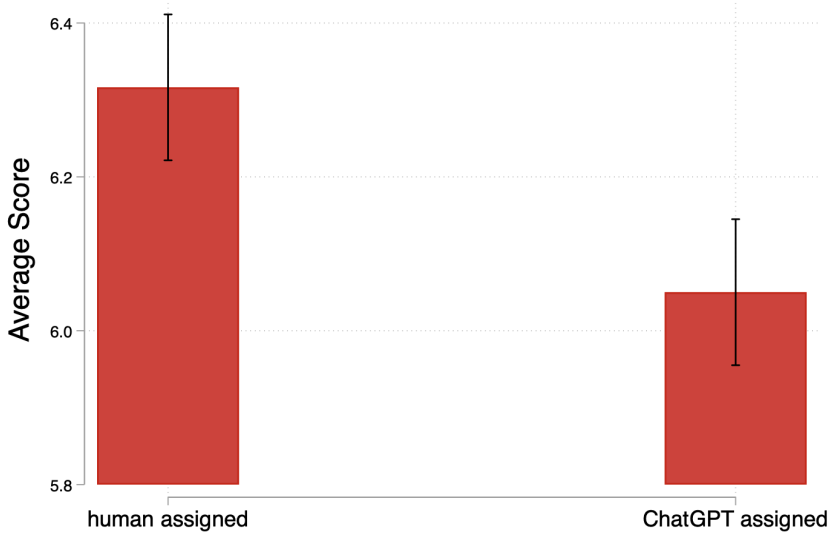
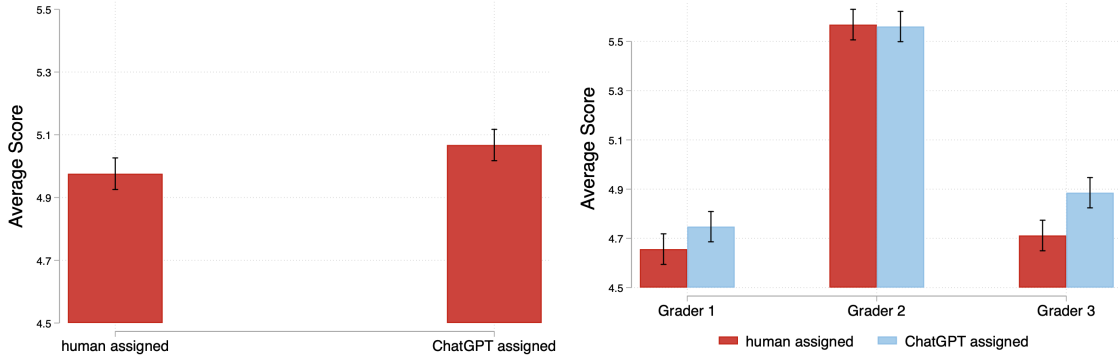


图 2：不同处理下的 ChatGPT评分。

为了生成由人类作出的评估，我们指示三名研究生使用提供给 ChatGPT 的相同提示词，对所有说明文字进行评分。图 3展示了评分结果，包括汇总结果以及每位评分者的评分明细。将三名人工评分者的评分取平均后，ChatGPT 评估处理组中的说明文字获得的平均评分在统计学上显著更高（5.07 对 4.97，双侧 t 检验，p 值 $= 0.0046$）。这一结果由三名评分者中的两名所驱动。



(a) 汇总所有评分者。

(b) 按评分者区分。

图3：不同处理下人工评估者的评分。

乍看之下，ChatGPT 和人工评分者给出的评分似乎有所不同，接下来的部分将更详细地探讨这些差异。

3.3 寻求外部帮助

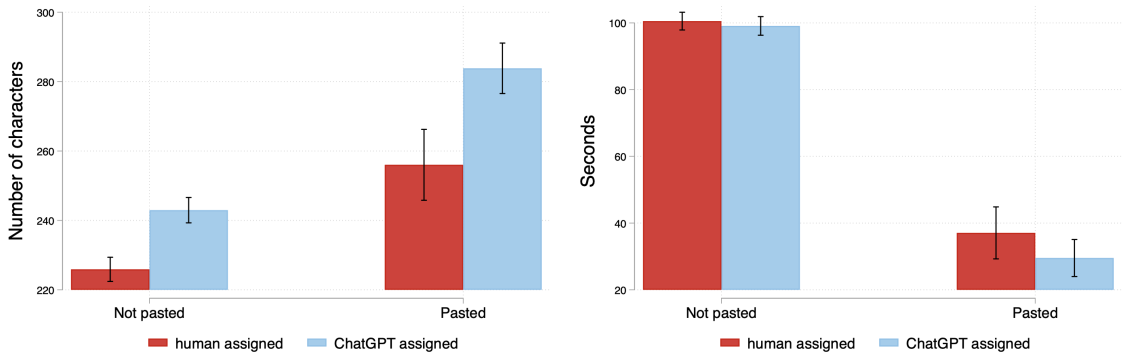
与此同时，人工智能在监控工作方面不断进步，也提升了执行受监控任务的能力，这使劳动者能够在日常工作中寻求人工智能的帮助。此前，我们报告称，被分配接受人工智能评估的参与者使用外部帮助的频率，是被分配接受人工评估的参与者的两倍（我们将粘贴虚拟变量作为外部帮助的代理指标）。这一发现与我们的预期一致：即使没有明确要求，人们也可能会对使用人工智能辅助（例如 ChatGPT）持负面评价。这也与 Awuah、Krenk 和 Yanagizawa-Drott（2024）同期开展的研究一致。该研究发现，非政府组织的申请评估者会对看起来由 ChatGPT 辅助撰写的求职信予以扣分，即使没有明确要求他们这样做。

本节将考察寻求帮助的两个方面。首先，我们分析观察到的写作习惯和质量方面的影响是否可归因于使用外部帮助。其次，我们比较收集到的多种外部帮助使用情况衡量指标，并介绍一项方法论贡献：一种新的引出方法，可用于识别在粘贴变量效果较差或无法使用的情境或任务中是否使用了外部帮助。

3.3.1 我们的结果是否由外部工具使用驱动?

通过分别分析包含粘贴文本和不包含粘贴文本的图像描述的写作习惯与质量，我们发现，图像描述会因是否粘贴文本而存在显著差异。然而，仅凭这些差异无法解释此前观察到的大部分处理效应。

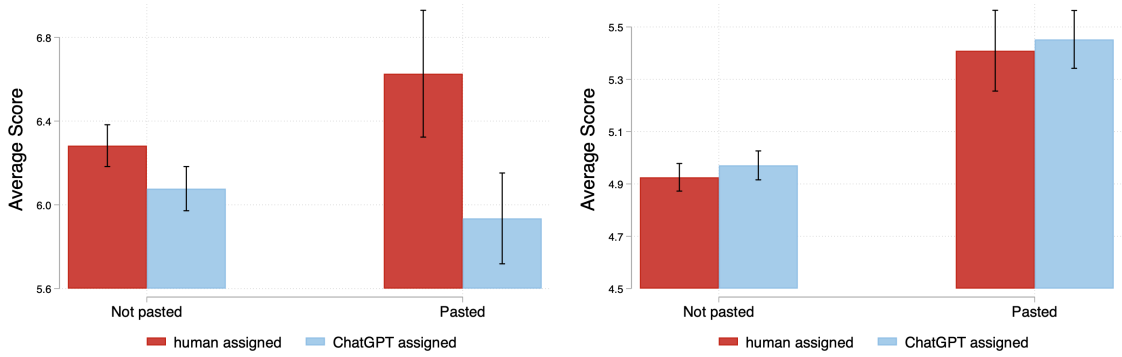
关于回复长度，我们观察到，粘贴的说明文字往往更长。不过，人工评估处理组的参与者平均写出的说明文字明显更短，无论这些说明文字是否由粘贴而来。就回复时间而言，未粘贴的说明文字耗时近三倍。控制粘贴因素后，回复时间不再存在处理效应。图 4展示了这些结果。



(a) 按粘贴划分的回复长度。 (b) 按粘贴划分的回复时间。

图4：按处理与粘贴划分的图像描述平均回复长度和时间。

重新考察说明文字的质量时，图5证实，无论是否发生粘贴，ChatGPT始终会给人工处理组的参与者评出更高的评分。相比之下，人工评分者给包含粘贴文本的说明文字评出的评分显著更高，这也解释了为何ChatGPT处理组参与者的平均评分更高，因为他们粘贴文本的频率是另一组的两倍。然而，分别分析包含和不包含粘贴文本的说明文字时，人工评分者给两个处理组评出的评分在统计上并无显著差异。



(a) 按粘贴划分的ChatGPT评分 (b) 按粘贴划分的人工评分

图5：按处理与粘贴划分的ChatGPT评分和人工评分。

3.3.2 衡量并引导外部帮助使用情况

本文使用文本粘贴作为识别外部协助的主要代理指标。然而，文本粘贴并非总是可行的方法（例如，对于非书面任务，或实施起来不切实际的情况）。我们提出了一种新的引出方法，

并检验其有效性，以扩展用于在实验中识别大语言模型使用情况的实验工具。

在实验结束时，我们会请参与者披露，他们是否使用过 ChatGPT 之类的工具来协助撰写任何描述。随后，我们向他们提供 1 美元奖金：如果他们能正确预测 ChatGPT 检测软件是否会将其任何描述判定为人工智能生成，便可获得这笔奖金。我们表明，为这个问题设置激励（如后一种情况）能带来更有希望的结果。

在非激励问题中，208 名参与者中有 13 人表示曾使用 ChatGPT 等工具获得外部帮助。在随后的激励问题中，承认自己的说明文字可能被认定为人工智能生成的参与者人数增加到了 29 人。有趣的是，最初在非激励问题中回答“是”的 13 名参与者中，有 3 人在激励问题加入检测工具后改答“否”。这 3 人中有 2 人在任何轮次都没有粘贴文本。这表明，尽管他们是自己撰写文本的，但可能在流程的其他环节使用了 ChatGPT。例如，一位参与者在开放式调查问题中提到：“有一张图片里，一位女士正在拍摄一个纸杯蛋糕。她撑着一把白色雨伞，所以我问 ChatGPT 这把伞是用来做什么的。”这个例子凸显了使用方式的多样性，既可能是完全委托，也可能只是收集信息以辅助工作。图6展示了参与者使用粘贴的轮次数分布，以及披露在引出问题中使用了帮助的参与者人数。结果表明，激励式方法特别善于识别频繁粘贴者（18–20 轮），检出了其中一半（26 人中的 13 人）。此外，激励问题相较于粘贴检测器还具有额外价值，识别出 13 名在任何轮次都没有粘贴的参与者。[16]

[16]如果将非激励问题中仅回答“是”的两位参与者也包括在内，这个数字就会增加到 15。

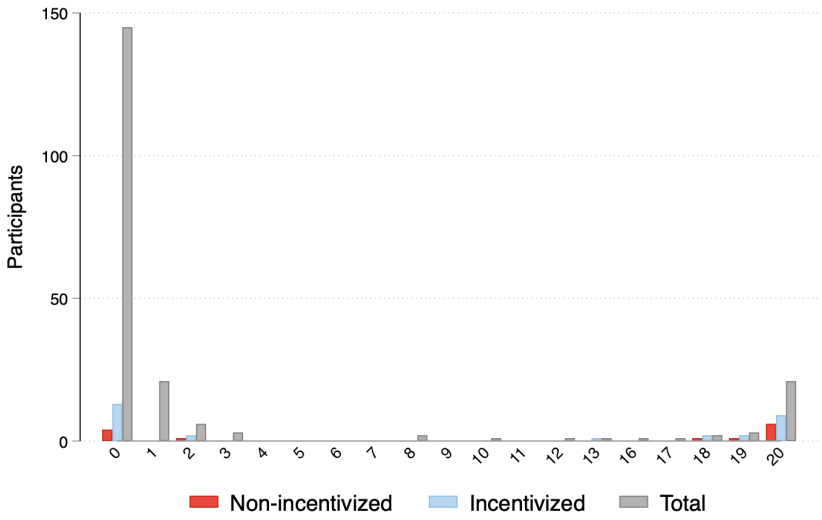


图6：参与者按粘贴文本的轮次数量分布。横轴表示参与者粘贴文本的轮次数量，红色柱表示在非激励问题中披露使用了帮助的参与者，蓝色柱表示在激励问题中披露使用了帮助的参与者，灰色柱表示各类别中的参与者总数。

4 讨论

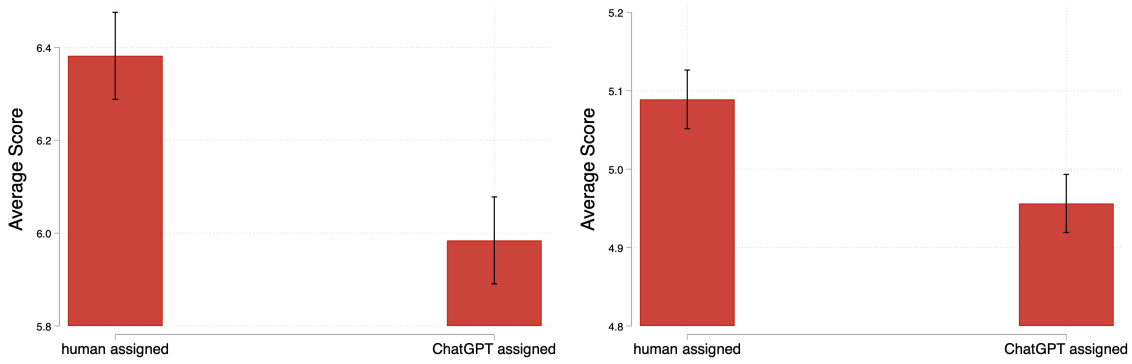
4.1 为什么处理效应只在 ChatGPT 评分中对质量产生影响?

在控制粘贴行为后，只有 ChatGPT 发现了不同处理组之间质量指标的显著差异。这引出了一个有趣的问题：为什么人工评分者没有像 ChatGPT 那样，给人工评估者处理组中的参与者更高的评分？由于说明文字的质量缺乏客观的真实值，两种评估来源不一定需要一致。不过，最初差异的一种可能解释是对文本长度的误解。人工评估者处理组中的参与者平均不仅写了更短的说明文字，而且如图A.2所示，文本长度还呈现一阶随机占优关系。

我们假设，参与者在撰写说明文字时，可能认为人工评分者会更偏好简洁（且可能更为精炼）的描述。然而，我们的研究结果表明，人工评分者给更长的说明文字打出了更高的评分。ChatGPT同样给更长的说明文字打出了更高的评分，但长度对评分的影响

对人工评分者更为显著。在将评分对处理和图注长度进行回归时，每增加100个字符，人工给出的评分估计提高1分，而对ChatGPT而言该效应为0.5分。这一结果并不令人意外，因为更长的说明文字往往体现出更大的努力并捕捉更多细节，因而可能产生更准确的描述。有趣的是，与ChatGPT相比，人工评估者更重视长度。

图7 表明，在控制文本长度后，ChatGPT和人工评估者均将人工评估处理中的作品质量评为显著更高。值得注意的是，即使不区分粘贴文本和非粘贴文本，这一结果仍然成立；而这种区分很可能会进一步放大处理效应。



(a) ChatGPT评分　　(b) 人工评分

图7：ChatGPT和人工评分 在控制文本长度后按处理划分 .

4.2 处理效应的解释

在考虑粘贴习惯后，我们发现不同处理组的说明文字长度存在差异。人工评估处理组的参与者写的说明文字更短，这表明他们在写作时可能更加谨慎。还有一项发现支持这一结论：控制文本长度后，无论是 ChatGPT 还是人类评分者，都会将那些为人工评分者写作的参与者所写的说明文字评为质量更高。

ChatGPT始终将人工评估处理中的说明文字评为更高质量。对于研究生，在控制文本长度时结果方向相同，而在其他情况下则不显著（当我们按粘贴进行区分时）。哪些因素可能解释为何被分配到人工评估的参与者似乎表现更好？为排除潜在解释，我们在各处理间固定了激励，发放相同数量的奖金。这确保关于评估者宽松度的相对信念无法解释

所观察到的差异。[17]

我们探讨了由人类而非人工智能评估者认可工作是否会带来不同的感受，这种感受促使被分配到人工评估组的参与者提高工作质量。我们询问参与者，当分配给他们的评估者给出高分时，无论他们获得多少报酬，他们会感到多开心（以 1 至 5 分计）。被分配给人工评估者的参与者在这一问题上的得分统计显著较高（4.41 对 4.09；双侧 t 检验，p 值 $= 0.0055$）。这表明，非金钱激励或许可以解释为什么被分配到人工评估组的参与者表现优于被分配到人工智能评估组的参与者。

5 结论

我们的研究提供了实证证据，说明当个人的工作由人工智能而非人类评估时，他们会如何改变行为。我们以图像描述这一常见工作任务为例，发现算法评估会增加产出数量。然而，在控制产出水平后，人工智能处理组的描述质量较低，无论评估者是人类还是人工智能。此外，人工智能评估会提高劳动者寻求外部协助的可能性，不过这本身无法完全解释质量下降。我们的结果表明，转向基于人工智能的评估系统可能会引发行为反应，企业和政策制定者在实施此类评估机制时应予以考虑。

这些发现对算法在组织中的未来应用具有重要意义。尽管人工智能评估具有成本效益且易于扩展，但其对工作质量的影响以及对外部工具的依赖，可能带来权衡。在质量至关重要的情境下，例如创意型和知识型任务，企业可能需要辅以人工审核，以减轻意料之外的后果。

未来的研究应探讨不同类型的人工智能评估如何影响工人在各类任务和行业中的行为。此外，人工智能评估对工人努力程度、学习和动机的长期影响仍有待探究。

[17]如果奖金是根据评估者是否认为图像描述合格来发放的，那么不同评估者对获得合格评分所需最低努力程度的看法可能有所不同，并可能影响参与者的努力程度。

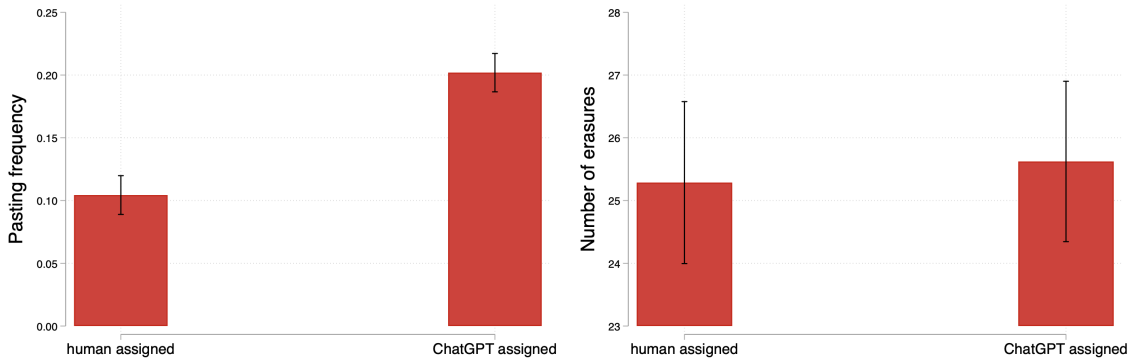
参考文献

Acemoglu, Daron 和 Pascual Restrepo（2019）。“自动化与新任务：技术如何取代并重新创造劳动”。《经济展望杂志》33.2，第 3–30 页。doi：10.1257/jep.33.2.3。Almog, David、Romain Gauriot、Lionel Page 和 Daniel Martin（2024）。“人类对人工智能监督的反应：来自中心球场的证据”。《ACM 人机交互论文集》。ACM 工作论文；DOI：10.1145/3670865.3673481。Autor, David、Caroline Chin、Anna Salomons 和 Bryan Seegmiller（2024）。“新前沿：新工作的起源与内容，1940–2018”。《经济学季刊》139.3，第 1399–1465 页。doi：10.1093/qje/qjae008。Awuah, Kobbina、Ursa Krenk 和 David Yanagizawa-Drott（2024）。“关于人工智能的学习”。工作论文。Bick, Alexander、Adam Blandin 和 David J. Deming（2024）。“生成式人工智能的迅速普及”。美国国家经济研究局工作论文 32966。2025 年 2 月修订。doi：10.3386/w32966。Brynjolfsson, Erik、Danielle Li 和 Lindsey Raymond（2025）。“工作中的生成式人工智能”。《经济学季刊》140.2。较早版本曾作为美国国家经济研究局工作论文 31161（2023）发表，第 889–942 页。doi：10.1086/705716。Camerer, Colin F.（2019）。“人工智能与行为经济学”。载于《人工智能经济学：一项议程》，第 587–608 页。Charness, Gary、Brian Jabarian 和 John A. List（2023）。“新一代：人工智能实验”。美国国家经济研究局工作论文 31679。doi：10.3386/w31679。Choi, Jonathan H. 和 Daniel Schwarcz（2025）。“法律分析中的人工智能辅助：一项实证研究”。《法律教育杂志》73。最初于 2023 年作为工作论文发布，第 384 页。Cohn, Alain、Tobias Gesche 和 Michel André Maréchal（2022）。“数字时代的诚实”。《管理科学》68.2，第 827–845 页。doi：10.1287/mnsc.2021.3985。Conlon, John J.（2025 年 6 月）。“记忆复述与信念偏差”。卡内基梅隆大学工作论文。首个版本：2024 年 5 月 28 日。本版本：2025 年 6 月 28 日。网址：https://johnjconlon17.github.io/website/conlon_rehearsal_beliefs.pdf。Corgnet, Brice、Roberto Hernán-González 和 Ricardo Mateo（2023）。“自动化世界中的同伴效应”。《劳动经济学》85，第 102455 页。doi：10.1016/j.labeco.2023.102455。Dickinson, David 和 Marie-Claire Villeval（2008）。“监控会降低工作努力吗？代理理论与挤出理论的互补性”。《博弈与经济行为》63.1，第 56–76 页。doi：10.1016/j.geb.2007.02.004。

Eloundou, Tyna、Sam Manning、Pamela Mishkin与Daniel Rock（2024）。“GPT即GPT：大语言模型对劳动力市场的潜在影响”。《科学》384.6702，第1306–1308页。doi: 10 .1126/science.adj0998。Filippas, Apostolos、John J. Horton与Benjamin S. Manning（2025）。“大语言模型作为模拟经济主体：我们能从Homo Silicus中学到什么？”工作论文。版本日期为2025年2月12日；较早版本以John J. Horton的NBER工作论文31122（2023）发表。Flodén, Jonas（2025）。“使用大语言模型进行考试评分：高等教育中人工评分与使用ChatGPT的人工智能评分之比较”。《英国教育研究期刊》51.1。在线优先发表：2024年9月16日，第201–224页。doi: 10.1002/berj.4069。url: https://doi.org/10.1002/berj.4069。Frey, Bruno（1993）。“监控会增加工作努力吗？信任与忠诚之间的竞争”。《经济探究》31.4，第663–670页。doi: 10.1111/j.1465-7295.1993.tb00897.x。Goergen, Jonas、Emanuel de Bellis与Anne-Kathrin Klesse（2025）。“人工智能评估改变人类行为”。《美国国家科学院院刊》122.25，e2425439122。doi: 10.1073/pnas.2425439122。Humlum, Anders与Emilie Vestergaard（2025）。“ChatGPT的不均衡采用加剧了劳动者之间既有的不平等”。《美国国家科学院院刊》122.1，e2414972121。doi: 10.1073/pnas.2414972121。Ide, Enrique与Eduard Talamas（2025）。“知识经济中的人工智能”。《政治经济学杂志》。2025年5月在线发表。doi: 10.1086/737233。Korinek, Anton（2023）。“面向经济学研究的生成式人工智能：用例及其对经济学家的启示”。《经济学文献杂志》61.4，第1281–1317页。doi: 10.1257/jel.20231736。Ludwig, Jens、Sendhil Mullainathan与Ashesh Rambachan（2024）。“算法的不合理有效性”。NBER工作论文32125。doi: 10.3386/w32125。March, Christoph（2021）。“人类与人工智能之间的策略互动：来自计算机玩家实验的启示”。《经济心理学杂志》87，第102426页。doi: 10.1016/j.joep.2021.102426。Mok, Lillio、Sasha Nanda与Ashton Anderson（2023）。“人们认为算法评估不如相同的人工评估公平和可信”。《ACM人机交互汇刊》7.CSCW2，第1–26页。Nagin, Daniel S.、James B. Rebitzer、Seth Sanders与Lowell J. Taylor（2002）。“监控、动机与管理：现场实验中机会主义行为的决定因素”。《美国经济评论》92.4，第850–873页。doi: 10.1257/00028280260344498。

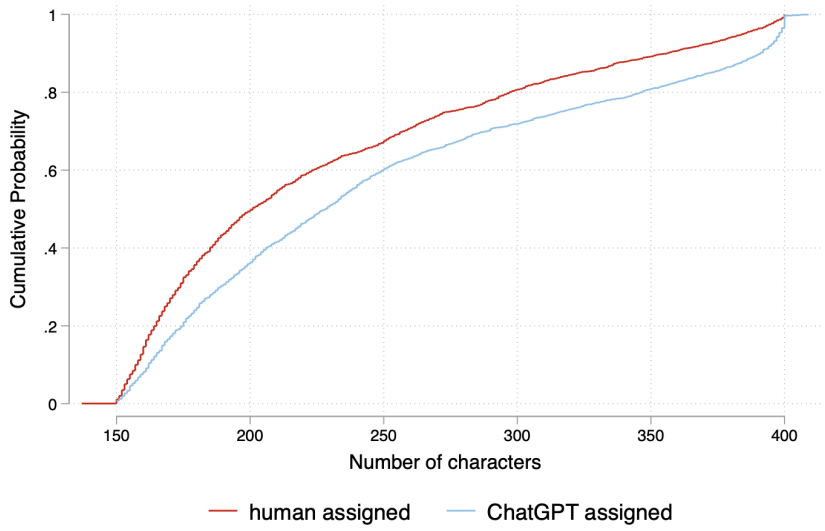
Noy, Shaked与Whitney Zhang（2023）。“关于生成式人工智能生产率效应的实验证据”。《科学》381.6654，第187–192页。doi: 10.1126/science.adh2586。Otis, Nicholas、Rowan Clarke、Solène Delecourt、David Holtz与Rembrand Koning（2025）。“生成式人工智能对创业绩效的不均衡影响”。开放科学框架预印本。预印本版本发布于2025年10月；针对肯尼亚企业家的现场实验。Palan, Stefan与Christian Schitter（2018）。“Prolific.ac——在线实验的受试者库”。《行为与实验金融杂志》17，第22–27页。doi: 10.1016/j.jbef.2017.12.004。Peng, Sida、Eirini Kalliamvakou、Peter Cihon与Mert Demirer（2023）。“人工智能对开发者生产力的影响：来自GitHub Copilot的证据”。arXiv预印本。doi:10.48550/arXiv.2302.06590。eprint: arXiv:2302.06590。Sarhan, Habiba与Simon Hegelich（2023）。“理解并评估政治图像中人工智能生成的图像说明文字的危害”。《政治科学前沿》5，第1245684页。doi: 10.3389/fpos.2023.1245684。Veselovsky, Veniamin、Manoel Horta Ribeiro、Philip Cozzolino、Andrew Gordon、David Rothschild与Robert West（2025）。“众包工作中大语言模型使用的普遍性与防范”。《美国计算机协会通讯》68.3。较早版本见于arXiv:2310.15683，第42–47页。Zhang, Da Wei、Melissa Boey、Yan Yu Tan与Alexis Hoh Sheng Jia（2024年12月）。“评估大语言模型用于基于标准的评分：从相符到一致”。《npj学习科学》9.1，第79页。doi: 10.1038/s41539-024-00291-1。Zhang, Simone、Janet Xu与AJ Alvero（2025）。“生成式人工智能遇上开放式调查回答：研究参与者对人工智能的使用与同质化”。《社会学方法与研究》54.3，第1197–1242页。doi: 10.1177/00491241251327130。

A 补充图



(a) 粘贴文本频率。 (b) 擦除。

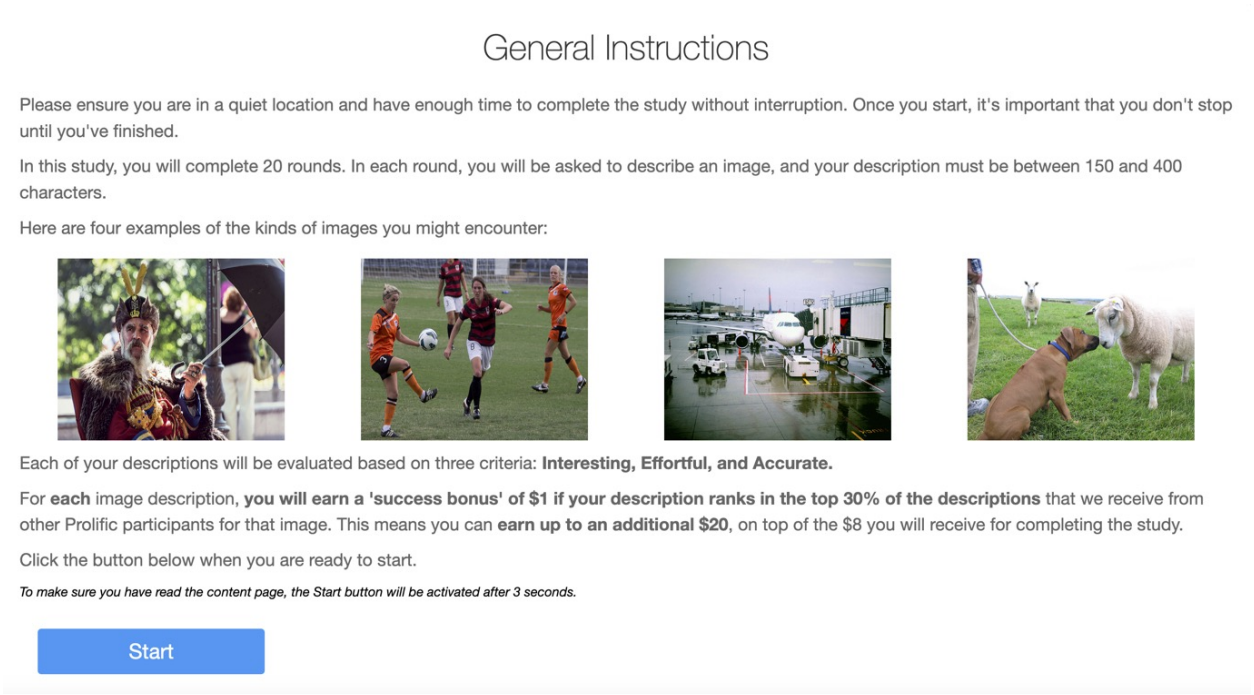
图 A.1：包含粘贴文本的说明文字频率，以及每条说明文字的平均擦除次数，按处理分组。

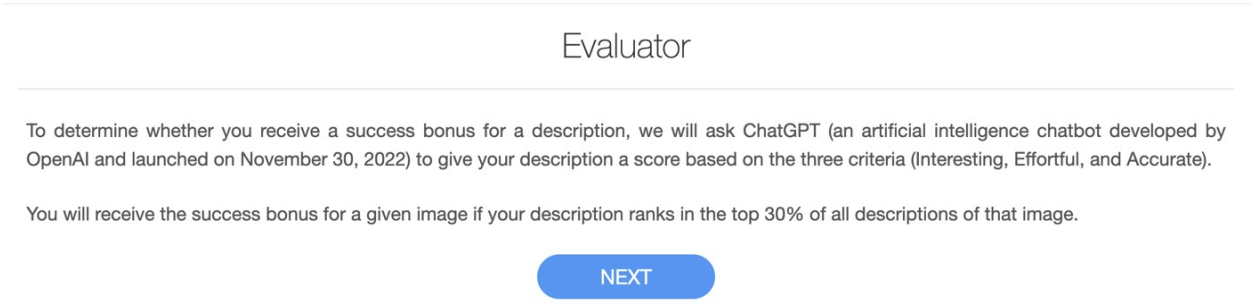


图A.2：按处理划分的图注长度经验累积分布函数。

B 实验设计

B.1 说明与界面





Round: 6 out of 20



How would you describe this image?

A muscular bulldog balances confidently on a bright blue surfboard, its sturdy frame adorned with a red lifesaver vest. The dog's tongue flops joyfully as it rides the crest of a shimmering wave, water spraying dramatically around it.

Character Count: 234

As a reminder, your descriptions will be evaluated by ChatGPT based on three criteria: Interesting, Effortful, and Accurate.

SUBMIT

We do not have a problem with you using a tool like ChatGPT, and the answer to these questions will not affect your compensation at all.

1. Did you use a tool like ChatGPT to help write any of your descriptions?

○ Yes ○ No

2. If you did use a tool like ChatGPT, please describe how you used it:

Next

You will earn an additional $1 bonus if you can correctly guess whether Originality.ai (a ChatGPT detector software) will say that any of your descriptions was written by ChatGPT. Again, we do not have a problem with you using a tool like ChatGPT, and the answer to this question will not decrease your compensation at all. In this case, you could even earn extra money.

Will Originality.ai (a ChatGPT detector software) say that any of your descriptions was written by ChatGPT?

○ Yes ○ No

Next

B.2 评估说明

我们向 ChatGPT 和负责评估的研究生提供了相同的说明。对于 ChatGPT，我们针对每条图像描述分别执行了以下三个提示词。

有趣
你将看到一张图像及其描述，并需要根据描述的有趣程度对其评分。

任务：评估描述的有趣程度，并给出 0 到 3 分。
评分标准：

- 0：不吸引人，也没有趣味。
- 1-2：比较吸引人，有一定趣味。
- 3：非常吸引人，富有创意。

步骤：
1. 阅读描述：仔细阅读提供的描述。2. 评估趣味性：判断描述是否吸引人、富有创意。3. 给予得分：根据上述标准为描述评分。

输出格式：
提供一个 单一得分（0–3），反映你认为该描述有多有趣。

用心
你将看到一幅图像及其描述，并需要根据描述所体现的用心程度对其进行评分。

任务：评估该描述的用心程度，并从 0 到 3 赋予得分。评分标准：

- 0：几乎没有付出努力。
- 1-2：表现出中等程度的努力和思考。
- 3：表现出高度的努力和周到的思考。

步骤：
1. 阅读描述：仔细查看所提供的描述。2. 评估努力程度：评估描述体现出的用心程度。3. 给出得分：根据上述标准为描述评分。

输出格式：
请给出一个得分（0–3），反映你认为这段描述有多用心。

准确
你将看到一张图像及其描述，并需要根据描述的准确程度为其评分。

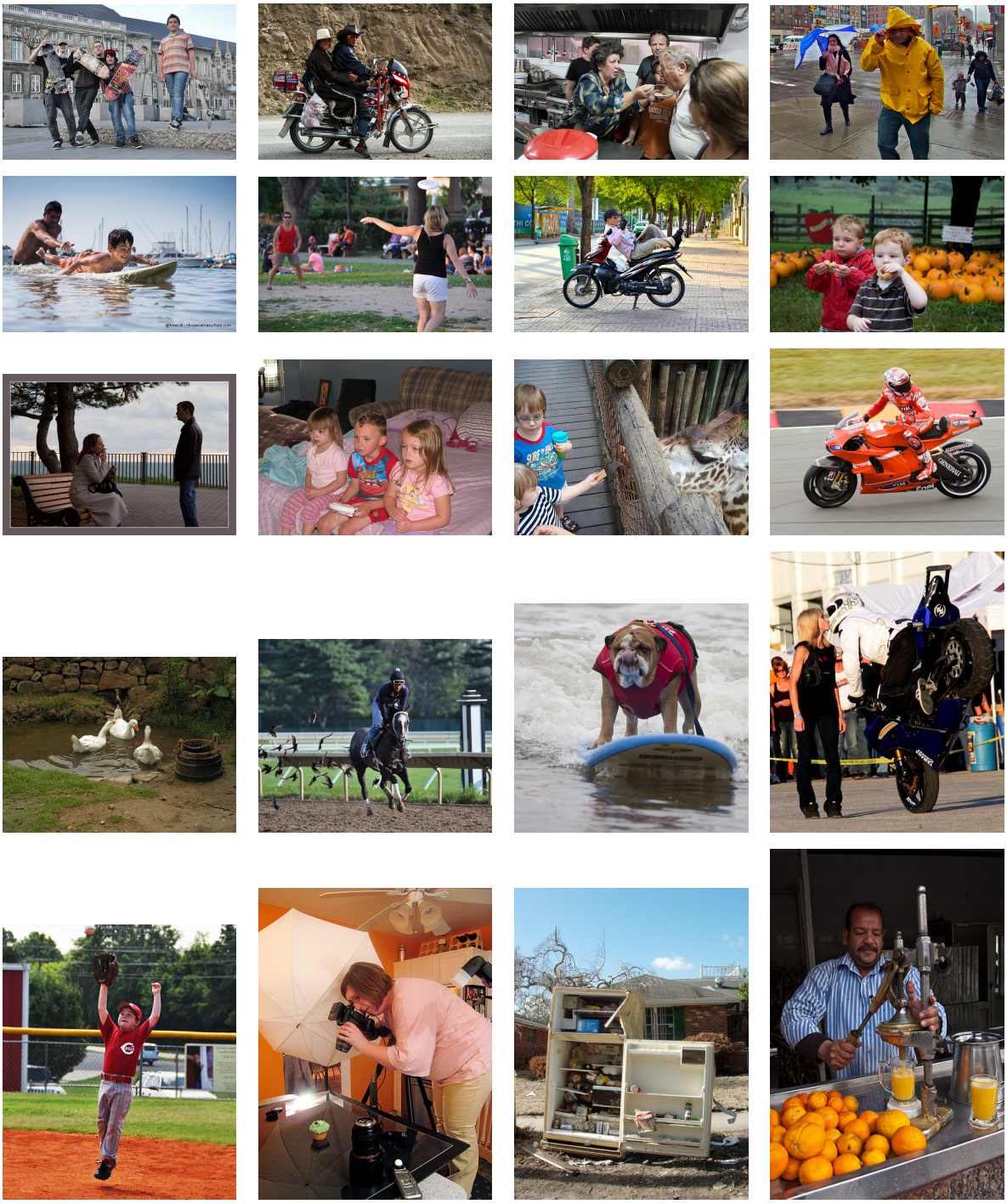
任务：评估描述的准确程度，并给出 0 到 3 的得分。评分标准：

- 0：与图像不符或无关。
- 1-2：基本准确，但存在一些小问题。
- 3：非常准确，与图像完全吻合。

步骤：
1. 阅读描述：仔细查看所提供的描述。2. 评估准确性：评估该描述给人的准确程度。3. 评定得分：根据上述标准为描述评分。

输出格式：
提供 一个 0–3 的得分，反映你认为该描述有多准确。

B.3 图像



C 试点结果

在第一个试点中，我们要求 ChatGPT 将说明文字评为满意或不满意。下文报告的结果根据所用标准是客观还是主观，以及这些标准是否以参与者最初选择的评估者为条件进行分类。

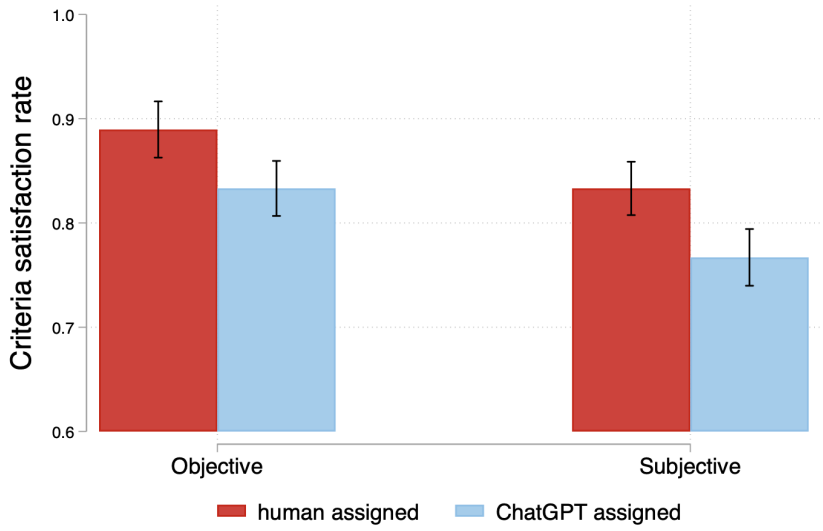


图 C.1：按处理和标准组划分的 ChatGPT评分。

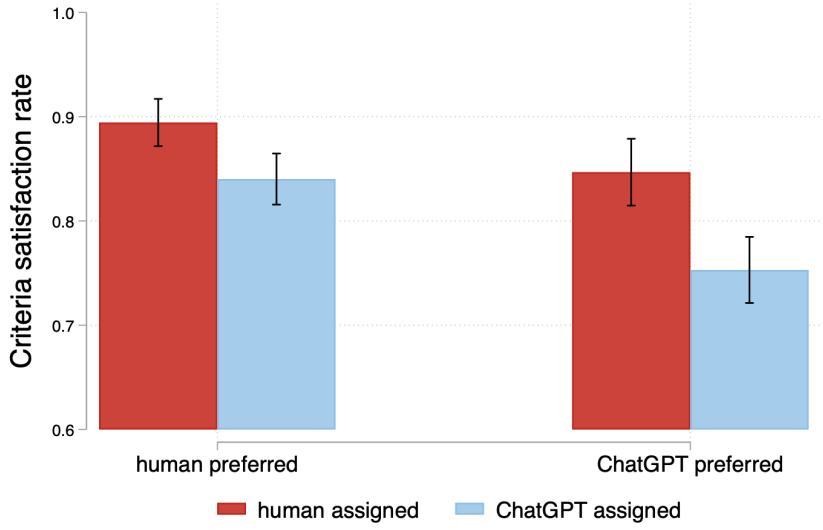


图 C.2：按处理和评估者偏好划分的 ChatGPT评分。