NOVARA
CONSULTING GROUP

This document was translated from English by machine and may contain mistakes. The English original is at https://www.novaracg.com/2025/05/23/four-numbers-and-no-counterfactual-reading-the-artificial-intelligence-case-study-in-human-resources/.

四个数字,没有反事实:解读人力资源领域的人工智能案例研究

Heather Grizzle · May 23, 2025

四项百分比提升,却没有反事实对照。每个标准人力资源*AI*案例研究指标要成立所需具备的前提,以及只需十分钟就能提出的问题。

目录

1. 这一体裁
2. 说法一:招聘周期缩短
3. 说法二:人工事务性工作量减少
4. 说法三:高潜力人才留任率提高
5. 说法四:消除偏见
6. 四者共同缺失的东西
7. 为何提出说法者往往真心相信
8. 买家应当提出的问题
9. 一份可信的案例研究应当包含什么
10. 结语
11. 参考文献

关于人力资源领域的人工智能案例研究已经形成了一种稳定的套路。

这一体裁

关于人力资源领域的人工智能案例研究已经形成了一种稳定的套路。某机构描述一次部署,报告百分之三或四的改善,加上一句关于消除偏见的说法,最后得出人才已成为竞争优势来源的结论。数字是具体的。方法论是缺失的。结论则邀人对话。

这种套路之所以延续,是因为它作为说服手段行之有效,而且没有人处于可以核实它的位置。供应商发布它,是因为它能促成销售。从业者发布它,是因为它能确立地位。买家接受它,是因为不接受就意味着承认自己没有依据做出即将做出的决定。链条中的每个人都对分子感兴趣,却没有人对分母感兴趣。

链条中的每个人都对分子感兴趣,却没有人对分母感兴趣。

Novara Consulting Group's position is that these documents should be read as advertising claims rather than as evidence, that the specific figures most often cited are the ones most easily produced without any underlying improvement, and that a buyer capable of asking four questions can distinguish a real result from a reported one in about ten minutes. What follows is that method, applied to the four claims that appear most frequently.

说法一:招聘周期缩短

招聘周期(time-to-fill)是最常被报告的指标,也是最容易在不改变招聘本身的情况下被“移动”的指标。

这项测量需要一个明确的起点和一个明确的终点,并在对比期间内保持不变。但在实践中,起点因机构而异,常常还因职位申请而异:它可能是职位申请被批准的日期、发布的日期、寻源开始的日期,或者用人经理首次提出需求的日期。将起点后移会缩短所测量的时间跨度,却不会缩短任何人实际经历的时间。排除被取消或撤回的职位申请、排除耗时最长的岗位,或改变分母中岗位的构成,效果同样如此。

混淆因素比效应本身更大。招聘周期对劳动力市场状况反应强烈,而当前大多数案例研究所涉及的对比期,恰恰跨越了从劳动力急剧短缺到明显宽松的转变。一个什么都没改变的机构,在这段时间跨度内也会显示出改善。要把这种变化归因于期间部署的某个系统,需要一个反事实场景,而这一体裁的案例研究中从来没有出现过这种东西。

限定语同样值得注意。仅针对“关键岗位”而非“所有岗位”报告的改善,表明这只是一个子集;而当这个子集是在结果已知之后才被挑选出来的,该数字所描述的其实是这次挑选本身,而非系统的作用。

说法二:人工事务性工作量减少

这一说法需要一个人工事务量的基线测量,而几乎没有哪个人力资源部门拥有这样的基线。

Where a figure exists it has usually been derived one of three ways: from a vendor′s implementation model, from a survey in which staff estimated their own time allocation, or from a count of tickets or transactions that changed definition when the new system was introduced. None of these supports a percentage comparison across time, because the instrument changed at the same moment as the intervention.

There is also a scope question the metric conceals. Transactional work that disappears from the human resources function frequently reappears elsewhere, most often with employees and line managers through self-service. A sixty percent reduction in transactions handled by the function may represent a genuine efficiency gain, a transfer of the same work to people whose time is not being counted, or some combination. Distinguishing these requires measuring total organizational effort, which nobody does, because the metric exists to justify the function′s investment rather than to describe the organization′s cost.

说法三:高潜力人才留任率提高

这是四项说法中最有意思的一项,因为产生这个数字的机制,往往就是部署本身。

“高潜力”不是一个可观察的特质,而是组织指定的一个标签,而人工智能人才系统通常会改变这一标签的指定方式。这正是它声称的目的。如果该系统改变了哪些员工被指定为高潜力人才,那么部署后被测量的群体,与部署前被测量的群体就是不同的人群,二者之间的留任率比较,并不是同一事物的比较。

由此产生的偏差方向是可以预见的。一个从绩效历史、敬业度信号和职业发展轨迹中识别高潜力的模型,会倾向于将那些敬业、近期获得晋升、正在上升的员工标记出来——而这些恰恰是预测员工会留下来的最有力指标之一。系统挑选出的是本来就打算留下的人,而组织却把他们的留任记录为系统的效果。

群体规模也加剧了这一问题。在大多数机构中,高潜力人才群体规模很小,以至于百分之十八的变化只代表寥寥数人,完全落在任何小群体年复一年自然波动的范围之内。这类数字理应附带群体规模,以及标签定义是否发生了变化的说明。但几乎从不这样做。

说法四:消除偏见

第四项说法与其他几项性质不同,因为它不仅仅是缺乏支撑,而是一种法律层面的陈述。

要证实这一说法,需要在明确界定的群体上,以足以支撑比较的样本量,对干预前后的偏见程度进行测量。这还要求这些群体是已知的,而这恰恰是该说法通常站不住脚之处。残疾这一维度根本无法用这种方式衡量:雇主在做出附条件录用之前,既不知道也不能合法询问哪些申请人存在残疾,因此任何影响比率工具都无法记录针对残疾的歧视。一个在继任规划中声称已消除偏见的机构,如果确实测量过什么的话,通常测量的是种族和性别,而对受绩效数据驱动模型影响最大的群体只字未提。

继任规划还带来另一重困难。一个从任职年限、职业阶段、在岗时间、资质获取时间以及此前晋升速度中识别“准备就绪程度”或“潜力”的模型,所依赖的信号与年龄高度相关。在没有审查过年龄分布的情况下,就为这样一个系统报告“消除偏见”,其实是在断言与输入变量所预示的恰恰相反的结论。

而这一说法本身就制造了风险敞口。经1991年修订的《民权法案》第七章规定,在与雇佣或晋升相关的候选人筛选或推荐中,基于种族、肤色、宗教、性别或民族出身调整分数、使用不同的录取分数线,或以其他方式改变与雇佣相关测试结果的做法属于违法。这里没有任何迹象表明这些系统存在此类行为。问题的关键在于:一份将自动化晋升工具与人口统计学结果相联系的公开声明,是该机构自己写下的文件,无法撤回,并且日后可能被要求做出解释。

四者共同缺失的东西

前述每一项失败,都是同一种缺失的不同表现。这些案例研究中没有一个包含反事实场景,也就是对“如果没有这次干预会发生什么”的描述。

该机构在劳动力市场发生转变、员工人数变化、经理人员更替、薪酬调整、以及其他举措同时进行的一段时期内部署了系统。随后,它将四项指标的变化归因于这些变化之一。这不是一个无关紧要的方法论小瑕疵,而是整个问题的核心,正是这一遗漏使得这一体裁无法被证伪。

关于相邻说法的已发表证据,应当让任何买家有所警觉。当哈佛商学院与Burning Glass Institute考察那些公开宣布采用技能导向招聘的公司实际发生了什么时,他们发现取消学历要求使非学历候选人的录用比例大约只提高了三点五个百分点,相当于每七百次录用中不到一次的变化,而且近半数公司尽管发表了声明,行为上却完全没有变化。这些声明是真诚的。实际做法与之完全脱节,而做出这些声明的机构自己也并不知情。

由此得出的合理推论并不是说人们在撒谎,而是说组织机构往往真的不知道自己的举措是否奏效,而案例研究恰恰是由最没有能力去弄清这一点的一方写出来的。

为何提出说法者往往真心相信

这一点值得明确说出来,因为另一种解读——即认为对方在故意作假——既刻薄又通常是错的。

这些数字很少是凭空捏造的。它们是在这样一种测量环境中产生的:没有建立基线,定义发生了变化,分析框架由供应商提供,而汇总结果的人正是当初倡导这笔采购的人。上述每一个条件都会把结果推向同一个方向,而这一切都不需要任何人怀有恶意。这个数字是“真实的”,在某种意义上说,确实有人计算出了它。只不过它并不代表其周围那句话所声称的含义。

这也是监管环境开始转变的原因。美国联邦贸易委员会已经在打击缺乏支撑的人工智能能力宣称方面采取了行动,包括2025年1月的一项命令,要求一家无障碍软件供应商就其自动化产品能使任何网站符合无障碍标准的宣称支付一百万美元。相关原则并不局限于该产品类别。关于人工智能性能的宣称需要有实证支撑,而一家机构在营销文件中公布结果百分比,正是在做出这样一种宣称。

买家应当提出的问题

四个问题,大约花十分钟即可问完,而绝大多数案例研究都经不起这样的追问。

基线是什么,何时测量的,由谁测量的。如果基线是在部署之后才构建出来的,或者是由供应商提供的,那么这个比较就根本不成立。

定义是否发生了变化。对于任何涉及指定群体的指标,比如“高潜力人才”或“关键岗位”,要问指定标准在前后是否一致。如果系统改变了这些标准,那么前后的群体就不同,比较也就失效了。

这段时期内还有什么其他变化。劳动力市场、员工人数、薪酬、管理层、同时进行的其他举措。诚实的回应者会列出一份清单。没有清单的回应者,是根本没有去核查过。

分母是什么,涉及多少人。没有群体规模的百分比是无法解读的,而小群体只需微小的变动就能产生很大的百分比。

> **第五个问题**
>
> 任何涉及偏见的说法都值得追问第五个问题:测量了哪些群体,样本量是多少,对于无法测量的群体又做了什么。最后这部分的答案几乎总是“什么都没做”,而这本身正是买家能获得的最有用的信息。

一份可信的案例研究应当包含什么

这一文体并非无可救药,任何愿意发布以下内容的机构都会立刻在一个无人这样做的领域中脱颖而出。

说明基线及其日期。说明定义,并注明是否有变化。报告队列规模及其对应的百分比。列出可能导致变化的同期其他因素,并说明哪些因素无法排除。描述哪些措施没有奏效,因为一次“一切都有所改善”的部署,恰恰说明什么都没有被真正测量过。若提出偏见方面的主张,须说明所涉群体、所用方法、样本量,以及尚未测量的部分。

这样的文件不如现在的形式那样光鲜亮丽,却实用得多。它也是唯一能经受住那些关心其真实性的人阅读的版本。

结语

案例研究是企业人力资源技术采购中最主要的证据材料,而在日常专业使用中,它也是最薄弱的证据形式。它来自单一机构、缺乏对照、自我报告、事后定义,并由有利害关系的一方发布。

None of that makes it worthless. It makes it a hypothesis. The organizations that will actually realize returns from these systems are the ones that treat other people's numbers as a reason to investigate rather than a reason to buy, and that hold their own numbers to the standard they would want applied to a vendor's. The discipline is

not difficult and it is not expensive. It is a baseline, a stable definition, a cohort size, and an honest account of what else was happening at the time.

参考文献

- 美国联邦贸易委员会,《关于*accessiBe, Inc.*案》,案号2223156,投诉与拟议裁决于2025年1月3日公布。
- Fuller, Joseph, 等. 《基于技能的招聘:从宣言到实践的漫长道路》. 哈佛商学院“管理未来工作项目”与Burning Glass Institute,2024年2月。
- 1964年《民权法案》第七章(经修订),42 U.S.C. § 2000e-2,含 § 2000e-2(l)。
- 《雇员选拔程序统一指南》,29 C.F.R. 第1607部分,适用于影响招聘、晋升及其他就业机会的选拔程序。

TIPS

Know something about a sign language AI system that the public should know? Send it to tips@novaracg.com. Every tip is read by me and kept confidential. If you work in the UK, writing to us is not by itself a protected disclosure under the Public Interest Disclosure Act 1998, so before you share employer information, check your options with Protect (protect-advice.org.uk), the free whistleblowing advice line.

引用方式

Grizzle, H. M. (2025, May 23). 四个数字,没有反事实:解读人力资源领域的人工智能案例研究. Novara Consulting Group. https://www.novaracg.com/zh/2025/05/23/four-numbers-and-no-counterfactual-reading-the-artificial-intelligence-case-study-in-human-resources/