

中国企业的人工智能投入： 来自招聘大数据的发现

方 颖 王翔宇 叶梦芊 赵西亮

人工智能是新一轮科技革命和产业变革的重要驱动力量。随着人工智能技术不断进步，越来越多企业在人工智能方面进行投入，以实现生产效率提升和生产方式转型升级。然而，人工智能技术投入难以通过财务报表等渠道直接观测，如何有效测算企业层面人工智能投入强度，成为人工智能相关研究亟待解决的重要课题。人工智能是典型的智力密集型技术，具备算法创新、数据处理能力并深度理解应用场景的人才，是人工智能投入的核心。基于此，本文提出一种基于企业招聘大数据的测算方法，以企业人工智能相关岗位招聘数量占总招聘数量的比例作为投入强度的代理指标，并通过招聘文本中的技能要求界定岗位属性。

现有测度方法主要包括文本法、设备投入法和专利法。文本法依赖企业披露文件中的相关词汇，但宣传强度与投入强度之间可能存在差异；设备投入法难以清晰区分人工智能专用设备和一般信息技术设备；专利法更多反映人工智能投入的潜在产出，而非投入的直接要素。与这些方法不同，本文从人力资本投入角度刻画企业人工智能投入，招聘数据更直接反映企业在特定时间和技术方向上的真实需求，信息密度更高，也更不易受到披露策略干扰。

本文构建了涵盖近 10 万项技能维度的中国劳动力技能大词典，并将其与大规模企业招聘数据进行匹配。具体做法是：从招聘文本中提取岗位技能集合，根据技能与人工智能核心技能共同出现的情况计算技能人工智能相关度，再在岗位层面判断岗位是否属于人工智能相关岗位，最后在企业年度层面计算人工智能相关岗位占全部岗位需求的比例。为验证测算方法的合理性，本文采用评级验证法、

专利验证法和资本市场验证法。结果显示，进入权威人工智能企业榜单的企业，其人工智能投入得分明显高于未上榜企业；企业人工智能投入水平越高，其人工智能专利申请量也相应增加。以 ChatGPT 发布作为人工智能技术进步的外生冲击后，高人工智能投入企业组合的超额收益率显著高于低投入企业组合，且人工智能投入与企业累计异常收益显著正相关。这表明本文构造的测度能够较好捕捉企业间人工智能投入强度差异。

基于这一测度，本文发现我国上市公司人工智能投入整体呈逐年上升趋势。从行业看，信息传输、软件和信息技术服务业以及金融业表现出最高的人工智能投入强度；农、林、牧、渔业，采矿业以及住宿和餐饮业投入程度较低。除少数行业外，其他行业 2019 年至 2022 年的投入强度相较于 2015 年至 2018 年均出现显著增长，说明人工智能技术正在逐步渗透到更多行业中。

本文进一步将“基于投入的测度”和“基于披露的测度”进行对比分析，目的在于揭示人工智能披露所真实承载的信息内容。研究发现，两种测度总体上呈正相关关系，但相关性较弱，说明存在较多投入与披露不一致的样本。将投入强度和披露强度转化为二元变量后，样本可分为“只说不做”“只做不说”“不说不做”和“言行一致”四类。结果显示，在所有至少进行过一次人工智能披露的样本中，“只说不做”的占比较高；在总样本中，也有较多企业既未投入也未披露。进一步分析表明，“只说不做”样本中人工智能披露对全要素生产率并未呈现显著促进作用，说明这类样本可能并未有效开展人工智能投入；但进一步分析显示，投入与披露不一致并不必然意味着粉饰，披露可能具有前瞻性，也可能受到企业业务与人工智能关联程度的影响。

企业的人工智能披露和投入信息会共同塑造投资者对于企业价值的判断，并最终表现为金融市场价格波动。因此，本文还考察投资者如何综合处理人工智能披露和投入两类信息。利用 ChatGPT 发布带来的外生事件窗口，本文计算各股票累计超额收益，并按高低

披露和高低投入形成四组。研究发现，无论投入高低，高披露相对低披露的收益优势始终显著，表明部分投资者主要依据披露信息评估企业人工智能水平；在高披露企业中，高投入相对低投入的溢价稳健存在，且“高投入高披露”股票表现最好。披露溢价在短期即显现，而投入溢价则在更长期限内逐步体现，说明投资者对披露信息的吸收更快，对投入信息的获取与理解存在时滞。

本文的边际贡献主要体现在三个方面：提出一种基于招聘大数据的方法来量化我国企业人工智能投入状况；采用“披露—投入”对照分析方法，系统考察我国企业人工智能信息披露的可靠性；通过考察外生技术冲击窗口内的资产表现，揭示我国投资者理解企业技术投入时的信息偏好，有助于厘清技术冲击从实体经济向金融市场传导的微观机制。

基于上述结论，本文提出以下政策建议。第一，推动传统产业人工智能应用转型。当前我国企业人工智能投入呈现显著行业分化特征，传统产业智能化投入整体相对不足，应坚持应用导向、场景牵引，降低企业智能化改造成本，并围绕质量检测、设备运维、供应链管理、客户服务等典型应用场景打造示范项目。第二，完善人工智能人才培养与流动机制，缓解企业投入中的人力资本约束，推动高校、职业教育与企业协同育人。第三，加强对人工智能扩散及企业分化的监测，防范“人工智能鸿沟”。第四，完善上市公司人工智能披露规范并强化真实性监管，关注招聘、研发和专利等可验证信息，强化一致性核查，减少“只说不做”带来的资源错配。

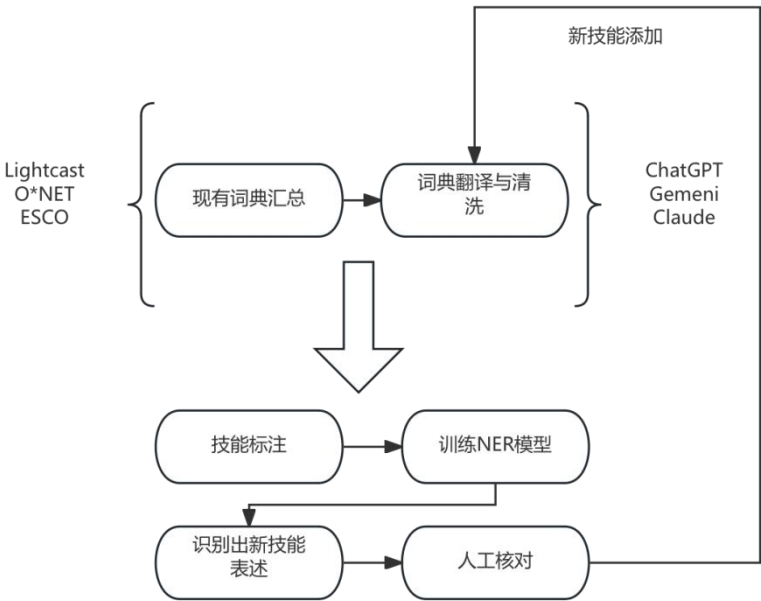
市场准入管制放松与企业技术多元化

王雄元 崔荡荡 王慧娴 王子平

市场准入负面清单制度通过逐步放松行政性管制、提升市场竞争的主导作用、重塑企业经营环境与竞争格局。竞争加剧不仅影响

《中国企业的人工智能投入：来自招聘大数据的发现》附录

附录 1 技能词典构建具体流程



附图 1 词典构建流程图

1. 初始英文技能库的整合与清洗

本文以 Lightcast、O*NET 和 ESCO 三套国际通行的英文技能数据库为基础。Lightcast 基于大量真实招聘信息构建，覆盖广、更新快，是招聘数据研究中常用的技能库，但相对更偏重具体技术技能。为补充通用技能与任务维度，本文进一步引入 O*NET 和 ESCO。对三套数据库进行清洗、标准化与去重后，得到初步英文技能词典。

2. 基于应用场景的英文技能翻译

O*NET 与 ESCO 的技能条目数量相对较少，且多为通用型技能，跨语境歧义较小，因此本文对其翻译结果逐条检查。相比之下，Lightcast 是词典主体，包含大量行业或领域特异性技能，仅依赖技能名称容易产生歧义。为提高翻译准确性，本文在翻译时同时引入 Lightcast 的 description 与 subcategory 作为上下文信息，要求模型结合应用场景完成区分。此外，本文以 ChatGPT 4o 为主要翻译器，并使用 Gemini 2.5 Flash 与 Claude Haiku 4.5 进行交叉验证。

3. 词典本土化扩展

仅依赖英文技能库翻译形成的中文词典，难以充分覆盖中文招聘语境中的缩写、行业惯用语和同义变体。为补充中国劳动力市场特有的技能表达，本文进一步在中文招聘文本中开展技能实体识别，采用基于 BERT (bert-base-chinese) 的命名实体识别方法。考虑到人工标注成本较高，本文使用“词典引导的弱监督标注”策略：先利用初版词典对招聘文本自动标注，再据此训练 NER 模型识别中文招聘语境中的技能实体，从而实现本土化扩展。

4. NER 模型效果的评估

本文在 2015—2022 年样本期内逐年随机抽取 10 万条招聘文本，形成共 80 万条文本的语料库。按常用“训练—验证—测试”划分原则，将语料随机划分为：10 万条训练集、10 万条验证集与 60 万条测试集（技能扩展集）。训练集与验证集均先使用初版技能词典进行自动

标注；在验证集上评估 BERT-NER 模型的技能识别性能，确认模型稳定性后，再将模型应用于 60 万条测试集以提取潜在新技能。

附表 1 汇报 BERT-NER 模型在 10 万条验证集上的技能提取表现：

附表 1 BERT-NER 模型表现评分

指标	中文名称	数值
<i>Precision</i>	精确率	0.9429
<i>Recall</i>	召回率	0.9533
<i>F1-score</i>	F1 值	0.9481
<i>Accuracy</i>	准确率	0.9912

结果显示模型在技能实体识别任务中具有较高精确率与召回率，且 F1 值接近 0.95，能够稳定地从中文招聘语境中识别技能表述，为后续本土化扩展提供了可靠基础。

5. 新技能筛选与扩展

将模型应用于 60 万条招聘文本后，共提取出 69,394 条不重复的潜在技能表述，其中 12,314 条已被初版词典覆盖。对其余 57,080 条未覆盖表述，本文仅保留在语料中出现次数超过 600 次、位于频率前 0.1% 的 401 项高频技能，以降低低频噪声干扰。新增词条中包括“PPT”“PS”“Office”等中文招聘语境中常见、但初版英文词典未充分覆盖的本土化表达。

综上，本文形成了一个较好兼顾覆盖范围与本土适配性的中国劳动力技能大词典，共包含技能词汇 99463 个。

附录 2 正文提及的图表

附表 2 2015 和 2022 年招聘大数据中 AI 关联性较高的技能

序号	技能	ω_s^{AI}
2015 年		
1	语义分析	0.7627
2	贝叶斯方法	0.7547
3	决策树	0.6718
4	语音识别	0.6666
5	逻辑回归	0.6612
6	推荐系统	0.6236
7	神经网络	0.5652
8	语音技术	0.5384
9	文本挖掘	0.5319
10	协同过滤	0.4929
2022 年		
1	自然语言理解	0.9577
2	文本分类	0.8955

3	Keras	0.8892
4	文本挖掘	0.8877
5	强化学习	0.8405
6	LSTM	0.8333
7	时间序列	0.8021
8	随机森林	0.7972
9	语音处理	0.7540
10	BERT	0.7526

附表 3 基于赛迪榜单的检验

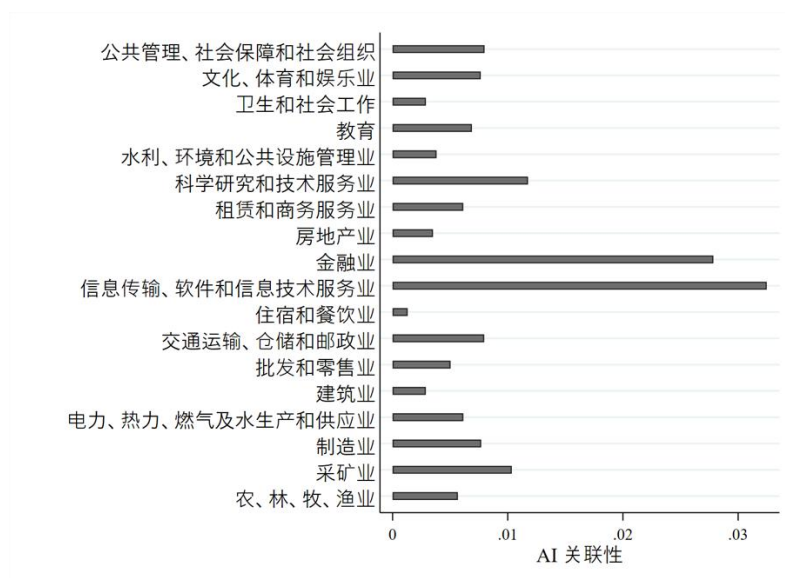
Panel A: 人工智能投入的描述性统计							
变量	样本	样本量	均值	标准差	最小值	中位数	最大值
<i>AIRatio</i>	全样本	26,029	0.0099	0.0351	0	0	1
<i>AIRatio</i>	上榜企业	103	0.0617	0.0854	0	0.036	0.5109
<i>AIRatio</i>	未上榜企业	25,926	0.0097	0.0360	0	0	1
Panel B: 人工智能投入的均值差异							
变量	样本	均值			差值		
<i>AIRatio</i>	上榜企业	0.0617			0.0521***		
<i>AIRatio</i>	未上榜企业	0.0097					

附表 4 企业人工智能投入与人工智能专利申请量的回归分析

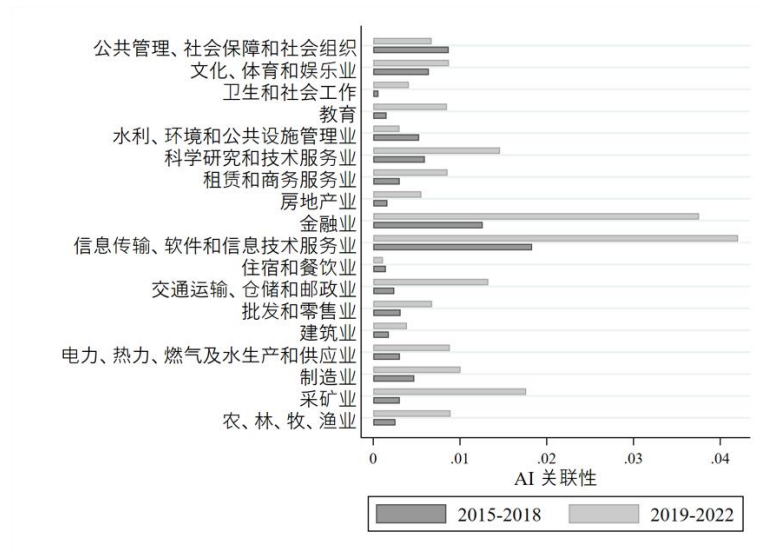
	(1)	(2)	(3)
变量	<i>lnAI</i>	<i>lnAI</i>	<i>lnAI</i>
<i>AIRatio</i>	2.9226*** (18.2746)	2.6080*** (18.9113)	2.5921*** (18.7453)
常数项	0.4603*** (6.7228)	0.5086*** (23.9899)	0.5111*** (24.0439)
行业固定效应	No	Yes	Yes
年份固定效应	No	No	Yes
样本量	12,894	12,894	12,894
<i>Adj. R</i> ²	0.4238	0.5000	0.5048

附表 5 人工智能投入与累计超额收益 CAR

变量	(1) CAR	(2) CAR	(3) CAR	(4) CAR
<i>AIRatio</i>	0.8918*** (9.1185)	0.3627*** (3.6431)	0.3568*** (3.5800)	0.3286*** (3.3195)
<i>Size</i>			0.0028 (1.2104)	0.0039 (1.4561)
<i>Big4Y</i>				0.0207* (1.7669)
<i>AnaAttention</i>				-0.0021*** (-5.5067)
<i>CompanyOpacity</i>				0.0228*** (4.9403)
常数项	-0.2428*** (-71.4587)	-0.1920*** (-4.0116)	-0.2023*** (-4.1611)	-0.2695*** (-5.4120)
行业固定效应	No	Yes	Yes	Yes
样本量	2,825	2,825	2,825	2,825
<i>Adj. R²</i>	0.0286	0.2288	0.2292	0.2418



附图 2 不同行业的人工智能投入（2015-2022 年）



附图 3 不同行业人工智能投入的动态变化（2015-2018 vs 2019-2022）

附表 6 言行分类情况的描述性统计

Types	Variable	Obs	Mean	Std. Dev.	Min	Max
只说不做	<i>LnwordsMDA</i>	5,548	1.4370	0.8139	0.6931	5.1818
	<i>AIRatio</i>	5,548	0	0	0	0
只做不说	<i>LnwordsMDA</i>	1,717	0	0	0	0
	<i>AIRatio</i>	1,717	0.0270	0.0577	0.0002	1
不说不做	<i>LnwordsMDA</i>	9,399	0	0	0	0
	<i>AIRatio</i>	9,399	0	0	0	0
言行一致	<i>LnwordsMDA</i>	3,097	1.9299	0.9964	0.6931	5.8021
	<i>AIRatio</i>	3,097	0.0316	0.0527	0.0003	1.0000