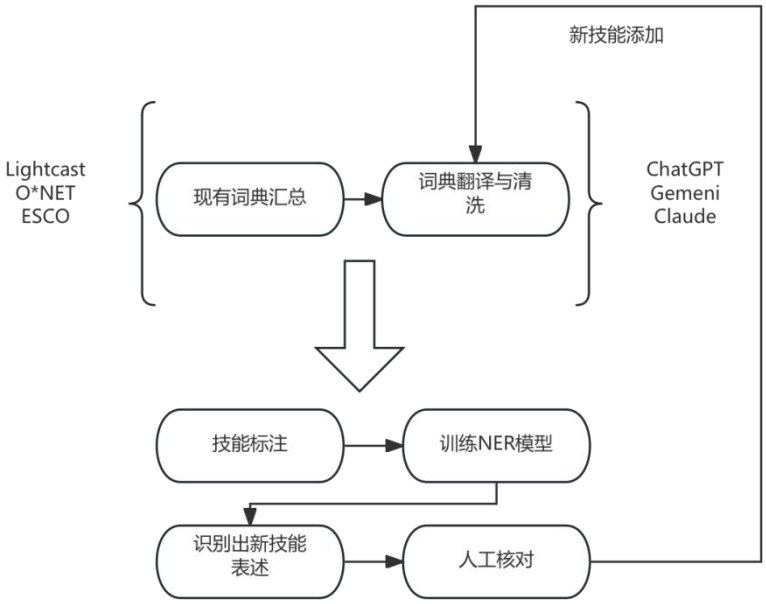
《中国企业的人工智能投入：来自招聘大数据的发现》附录

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
Precision	精确率	0.9429
Recall	召回率	0.9533
F1-score	F1 值	0.9481
Accuracy	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
AIRatio	全样本	26,029	0.0099	0.0351	0	0	1
AIRatio	上榜企业	103	0.0617	0.0854	0	0.036	0.5109
AIRatio	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

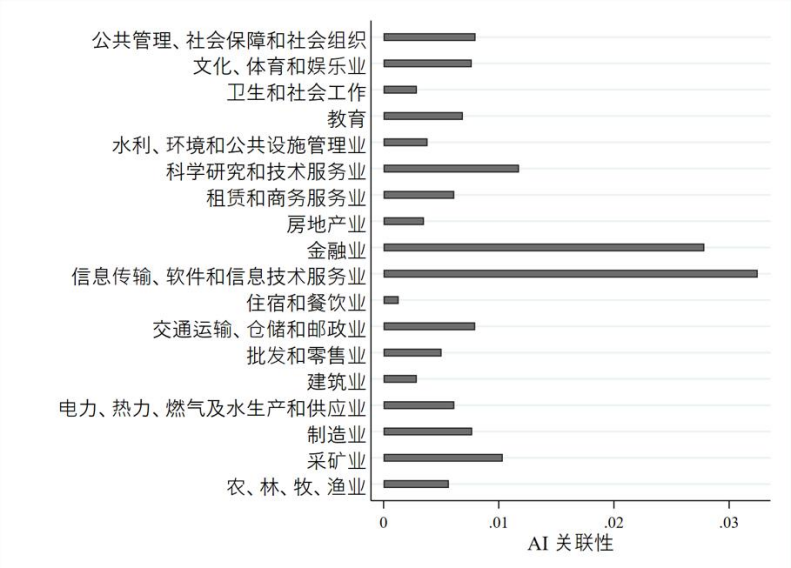
变量	样本	均值	差值
AIRatio	上榜企业	0.0617	0.0521***
AIRatio	未上榜企业	0.0097	

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

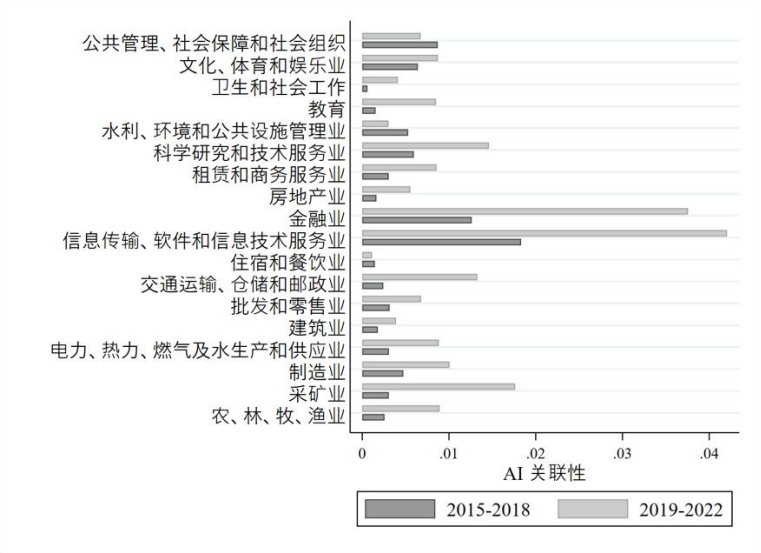
	(1)	**(2)**	**(3)**
变量	*lnAI*	*lnAI*	*lnAI*
AIRatio	2.9226***	2.6080***	2.5921***
	(18.2746)	(18.9113)	(18.7453)
常数项	0.4603***	0.5086***	0.5111***
	(6.7228)	(23.9899)	(24.0439)
行业固定效应	No	Yes	Yes
年份固定效应	No	No	Yes
样本量	12,894	12,894	12,894
Adj. R^2	0.4238	0.5000	0.5048

附表 5 人工智能投入与累计超额收益 CAR

	(1)	(2)	(3)	(4)
变量	*CAR*	*CAR*	*CAR*	*CAR*
AIRatio	0.8918***	0.3627***	0.3568***	0.3286***
	(9.1185)	(3.6431)	(3.5800)	(3.3195)
Size			0.0028	0.0039
			(1.2104)	(1.4561)
Big4Y				0.0207*
				(1.7669)
AnaAttention				-0.0021***
				(-5.5067)
CompanyOpacity				0.0228***
				(4.9403)
常数项	-0.2428***	-0.1920***	-0.2023***	-0.2695***
	(-71.4587)	(-4.0116)	(-4.1611)	(-5.4120)
行业固定效应	No	Yes	Yes	Yes
样本量	2,825	2,825	2,825	2,825
Adj. R^2	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	*LnwordsMDA*	5,548	1.4370	0.8139	0.6931	5.1818
	AIRatio	5,548	0	0	0	0
只做不说	*LnwordsMDA*	1,717	0	0	0	0
	AIRatio	1,717	0.0270	0.0577	0.0002	1
不说不做	*LnwordsMDA*	9,399	0	0	0	0
	AIRatio	9,399	0	0	0	0
言行一致	*LnwordsMDA*	3,097	1.9299	0.9964	0.6931	5.8021
	AIRatio	3,097	0.0316	0.0527	0.0003	1.0000