第 1 章

数据标注概述

无人驾驶、人脸识别、语音交互……在人工智能（artificial intelligence，AI）第三次浪潮之下，在算力、算法与数据的合力推动下，人工智能技术突破与行业落地如雨后春笋，焕发出源源不断的生机。尤为瞩目的是，在灼热的人工智能发展背后，为其发展提供数据燃料的数据标注正在成为一门新兴产业。

与此同时，数字经济的蓬勃发展，也为数据标注的发展进一步助燃。根据中国信息通信研究院的数据，我国数字经济的规模从 2005 年的 2.6 万亿增长到 2020 年的 39.2 万亿。到 2025 年，数字经济带动就业人数将达到 3.79 亿[1]。数字经济的不断发展正在催生更多新职业，人工智能训练师就是其中之一。

1.1 数据标注的起源与发展

由于数据标注与人工智能相伴相生，在研究数据标注的同时，首先需要对人工智能追本溯源。人工智能的概念最早由约翰·麦卡锡于 1956 年达特茅斯会议上提出，意指让机器具有像人一般的智能行为。

在其提出以来的 60 多年中，人工智能的发展并非坦途，而是经历了沉沉浮浮、三起三落。人工智能在达特茅斯会议上经过了两个多月的讨论，并在会后推出了第一款聊天软件，让人直呼“人工智能来了”，并掀起了此后为期 20 年的第一次人工智能浪潮。

当时主要以注重演算与推理的符号主义以及深度学习的“前身”——连接主义为代表。对于此次人工智能的兴起，专家学者尤为看好，甚至指

出，未来十年机器人就能够超越人类了。然而，就在大家期盼人工智能春天到来之际，在 20 世纪 70 年代后期，人们却逐渐发现过去的理论与模型只能用于解决一些简单问题，同时运算能力不足，人工智能的第一次浪潮偃旗息鼓，进入了突如其来的冬天。

此后，经过短暂的消沉后，随着 20 世纪 80 年代两层神经元网络（BP 网络）的兴起，人工智能开始焕发出新的生机，迎来了第二次发展浪潮。其间，语音识别、语音翻译以及感知机模式成了典型代表。然而，这些现在看来再寻常不过的应用，彼时离人们的实际生活仍然较为遥远，人工智能也随之进入了第二次沉寂的低潮，人工智能发展历史如图 1-1 所示。

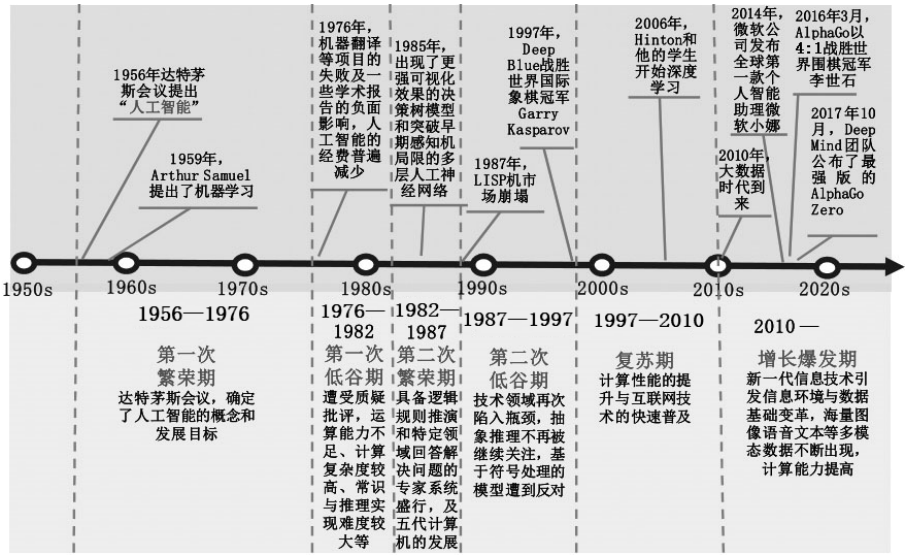


图 1-1　人工智能发展历史[2]

人工智能的第三次浪潮始于 Deep Blue（IBM 深蓝）的出现，其在 1997 年战胜了国际象棋冠军，而 2006 年“神经网络之父”杰弗里 • 辛顿（Geoffrey Hinton）提出的深度学习技术进一步助推人工智能的发展，该技术于 2010 年大火，直接带动了人工智能的真正爆发，使其成为了商界、创投界炙手可热的新星，并发展至今。不难预见，未来人工智能将实现由弱人工智能到强人工智能，直至超人工智能的高度。

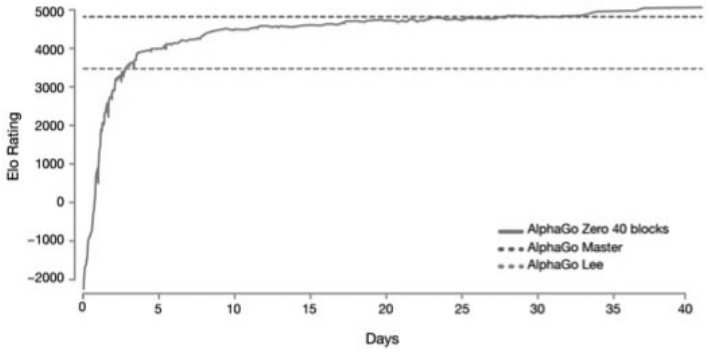
纵览人工智能的发展脉络，在前两次发展浪潮中，人工智能发展起起伏伏，偶有爆发，但却未能真正深入人们的生活。因此，当时由于量级比较小，为人工智能提供“喂养数据”的数据标注主要由研究该领域的工程师完成，并不能称之为独立的职业。近年来，随着人工智能第三次浪潮的到来，数据标注的需求应接不暇，2011 年数据标注的外包市场开启，2017 年真正爆发，数据标注开始慢慢进入人们的视野。

《“十四五”数字经济发展规划》明确指出，“充分发挥数据要素作用”

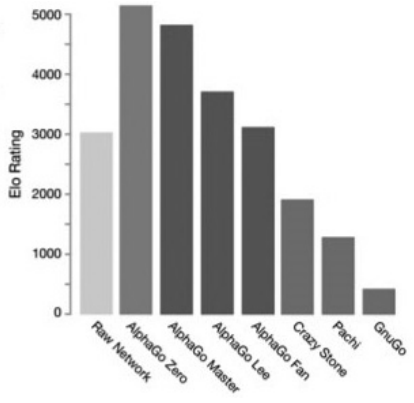
“强化高质量数据要素供给。支持市场主体依法合规开展数据采集，聚焦数据的标注、清洗、脱敏、脱密、聚合、分析等环节，提升数据资源处理能力，培育壮大数据服务产业”，并在数据要素供给、数据要素市场化、数据要素开发利用机制等方面进行了部署[3]。据 iResearch 数据显示，预计 2025 年数据标注行业市场规模将突破 100 亿元[4]。

1.1.1　什么是数据标注

2016 年，人工智能程序阿尔法围棋（AlphaGo）在与世界顶尖棋手的对决中奉上了令人惊艳的战绩，一战成名。此后横空出世的阿尔法零（AlphaGo Zero）作为 AlphaGo 的升级版本，自学 3 天，以 100∶0 的成绩完胜此前击败李世石的 AlphaGo 版本；自学 40 天，以 89∶11 的绝对优势击败阿尔法大师（AlphaGo Master）版，不同 AlphaGo 版本的棋力比较如图 1-2 所示。



（a）



（b）

图 1-2　不同 AlphaGo 版本的棋力比较[5]

当我们感慨其成长速度时，不得不承认，最初的 AlphaGo 也犹如出生的婴儿一般，对下棋一窍不通，其之所以能够快速升级成为棋坛高手，是

因为人类“喂养”的棋谱与数据，换言之，正是人类像教育小孩一样培养了 AlphaGo，才让其“学会”下棋。

举个简单的例子，我们告诉孩子——“这是一辆汽车”，并把对应的图片展示在孩子面前，帮助他记住了拥有四个轮子，可以有不同颜色的这种日常交通工具，当孩子下次在大街上遇到飞奔的汽车时，也能直呼“汽车”。

类比机器学习，如果准备让机器习得同样的认知能力，我们也需要帮助机器识得相应特征。两者不同点在于，对于人类来说，往往告诉他一次就能记住，下次遇到就能准确辨别；对于机器来说，需要我们提取有关汽车的特征，“喂”给他们大量带有汽车特征的图片，使其通过训练集反复学习，通过测试集进行检查与巩固，最终能够准确识别汽车，而这些带有汽车特征的图片正是出自数据标注。

简而言之，数据标注即通过分类、画框、标注、注释等，对图片、语音、文本、视频等数据进行处理，标记对象的特征，以作为机器学习基础素材的过程。由于机器学习需要反复学习以训练模型和提高精度，同时无人驾驶、智慧医疗、语音交互等各大应用场景都需要标注数据，人工智能训练师应运而生。

1.1.2 数据标注分类概述

对于数据标注，按照不同的分类标准，可以有不同划分。下面以标注对象作为分类基础，将数据标注划分为图像标注、语音标注、文本标注以及视频标注。

1．图像标注

提及数据标注，大多数人第一反应就是图像标注。图像标注是一个将标签添加到图像上的过程。图像标注类型包括拉框、语义分割、实例分割、目标检测、图像分类、关键点、线段标注、文字识别转写、点云标注、属性判断等。图像标注在人工智能与各行各业应用相结合的研究过程中扮演着重要的角色：通过对路况图片中的汽车和行人进行筛选、分类、标框，可以提高安防摄像头以及无人驾驶系统的识别能力，如图 1-3 所示；通过对医疗影像进行骨骼描点，特别是对病理切片进行标注分析，能够帮助 AI 提前预测各种疾病。

2．语音标注

语音标注是把语音中包含的文字信息、各种声音“提取”出来，再进行转写或者合成，从而用作人工智能机器学习数据。语音标注类型包括 ASR 语音转写、语音切割、语音清洗、情绪判定、声纹识别、音素标注、韵律标注、发音校对等。目前，在人工智能研究中，语音应答交互系统是一个重要分支，其中聊天机器人人气颇高，苹果的 Siri、小米的小爱同学等已经

成为深入日常生活的重要应用。在此类虚拟助理的研发过程中，基于语音识别、声纹识别、语音合成等建模与测试需要，需要对数据进行发音人角色标注、环境情景标注、多语种标注、ToBI（tones and break indices）韵律标注体系标注、噪音标注等，如图 1-4 所示。

图 1-3　图像标注[6]

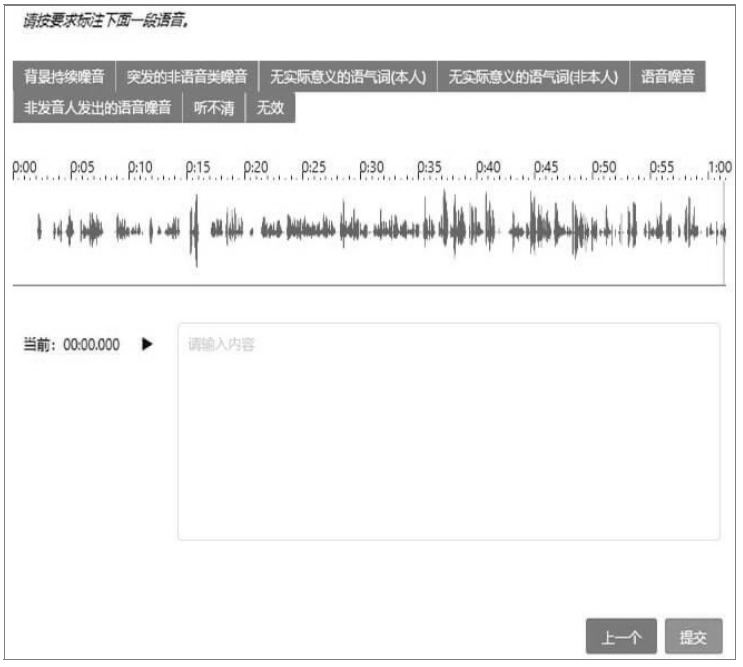


图 1-4　语音标注

3．文本标注

文本标注是对文本进行特征标记，为其打上具体的语义、构成、语境、

目的、情感等原数据标签，主要用于自然语言处理。自然语言处理是人工智能的分支学科，在满足自然语言处理不同层次需要的过程中，对文本数据进行标注处理是关键环节。具体而言，通过语句分词标注、语义判定标注、文本翻译标注、情感色彩标注、拼音标注、多音字标注、数字符号标注等，可获得高准确率的文本语料，如图 1-5 所示。

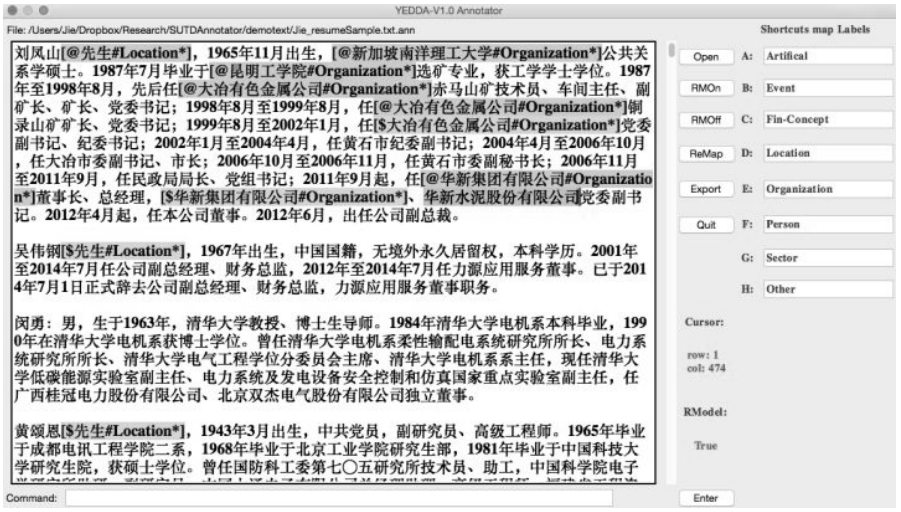


图 1-5　文本标注

4．视频标注

视频标注以图片帧为单位，对视频素材中的目标对象进行跟踪，对包括道路、车辆、行人等在内的目标物的特征信息、结构信息、语义信息等进行标记，从而形成训练数据集。与图像标注相比，视频标注不只限于一张图片，而是对某段时间内连续的一系列图像数据进行标记和汇总，生成的内容丰富而直观。按照具体应用类型，视频标注可进一步划分为视频追踪、标签分类、视频打点以及视频信息提取，如图 1-6 所示。

图 1-6　视频标注

1.1.3　数据标注流程概述

数据标注的质量直接关系到模型训练的优劣程度，因此数据标注需要建立一套既定的数据标注流程，对图像、语音、文本、视频等进行有序而有效的标注，如图 1-7 所示。

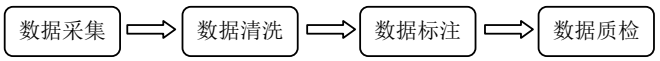


图 1-7　数据标注流程

1．数据采集

数据采集与获取是整个数据标注流程的首要环节。目前对于数据标注众包平台而言，其数据主要源于提出标注需求的人工智能企业。对于这些人工智能企业，他们的数据又来自哪里？比较常见的是通过互联网获取公开数据集与专业数据集。公开数据集是政府、科研机构等对外开放的资源，获取比较简便，而专业数据集的获取往往更耗费人力物力，有时通过购买所得，有时通过拍摄、截屏等方式积累素材再自主整理所得。此外，对于 Google 等科技巨头而言，其本身就是一个巨大的数据资源库。

至于具体的数据获取方式，既可以通过内部数据库，以 SQL 技能提取数据，也可以下载获取政府、科研机构、企业开放的公开数据集。此外，还可编写网页爬虫，收集互联网上多种多样的数据，比如爬取知乎、豆瓣、网易等网站的相关数据。

值得一提的是，在进行数据采集时，不仅需要考虑采集规模与预算，还应注重采集数据的多样性以及数据对应用场景的适用性。再者，数据采集应该合法合理，通过正当的方式获取，不能侵犯他人隐私权、肖像权等个人权利，这是数据采集的前提。

2．数据清洗

在获取数据后，并不是每一条数据都能够直接使用，不少数据是不完整、不一致、有噪声的脏数据，需要通过数据预处理，才能真正投入问题的分析研究中。在预处理的过程中，旨在把脏数据“洗掉”的数据清洗是重要一环。

在数据清洗中，应对所有采集的数据特别是一些爬虫数据以及视频监控数据进行筛检，去掉重复的、无关的内容，对异常值与缺失值进行查漏补缺，同时平滑噪声数据，最大限度纠正数据的不一致、不完整，将数据统一成适合标注且与主体密切相关的标准格式，以帮助训练更为精确的数据模型和算法。

3．数据标注

数据经过清洗，即可进入数据标注的核心环节。一般在正式标注前，

会由需求方的算法工程师给出标注样板，并为具体标注人员详细阐述标注需求与标注规则，经过充分讨论与沟通，以保证最终数据输出的方式、格式以及质量一步到位，这也被称为试标过程。

试标后，标注工程师按照此前沟通确认的要求进行数据标注，通过对图像、视频、语音、文本等素材进行细致的分类、标框、描点等操作，给素材打上不同的标签，以满足不同的人工智能应用需要。

4．数据质检

无论是数据采集、数据清洗，还是数据标注，人工处理数据的方式并不能保证这些过程完全准确。为了提高输出数据的准确率，数据质检成为了重要一环，而最终通过质检环节的数据才算是真正过关。

对于具体质检而言，可以通过排查或抽查的方式。检查时，一般设有多名专职的审核员，对数据质量进行层层把关，一旦发现提交的数据不合格，直接交由数据标注人员返工，直至最终通过审核为止。

1.2 数据标注的应用案例

无论是全职还是兼职，数据标注人员数量之所以创新高，主要归因于呈现指数级增长的人工智能发展，以及随之而来的日趋多样化的数据标注应用场景。

1.2.1 出行行业

对于出行行业而言，数据标注除了用于汽车自动驾驶研发之外，结合物联网数据、交通网络大数据以及车载应用技术，则能进一步帮助规划出行路线，优化驾驶环境。以下是数据标注常见应用：以矩形框或描点对车辆进行标注，标记拍到的物体是活体、障碍物还是其他物体；以矩形框或描点标注人体轮廓；采集地址兴趣点，在地图上做出相应地理位置信息标记的 POI（point of interest）标记等。

例如，在自动驾驶领域，Scale 公司目前通过提供图像标注、图像转录、分类、比较和数据收集的 API（应用程序界面），以目标识别来标注数据集。具体而言，在传感器与 API 的融合应用下，通过对相机、激光雷达和 Radar 数据进行标记，对周围环境状况，包括汽车与其他物体的距离、移动速度等进行标注，生成可用于训练 3D 感知模型的标注数据[7]。

1.2.2 金融行业

目前，人工智能的触角逐渐伸向金融领域。无论是身份验证、智能投资顾问，还是风险管理、欺诈检测等，以高质量的标注数据提高金融机构

的执行效率与准确率，已经成为一大趋势。其中，文字翻译、语义分析、语音转录、图像标注等，都是具有代表性的重要应用。

一直以来，对于金融合同而言，往往需要花费律师或贷款人员大量时间进行核对与确认。摩根大通开发了一款 AI 软件，通过语义分析处理的数据训练，使得原来需要 36 万个小时完成的合同审查工作数秒即可完成，而且错误率大大降低。

1.2.3　医疗行业

在医疗行业，通过人体标框、3D 画框、骨骼点标记、病历转录等应用，机器学习能够快速完成医学编码和注释，以及在远程医疗、医疗机器人、医疗影像、药物挖掘等场景的应用，助力提供更高效的诊断与治疗，制订更为健全的医疗保险计划。

比如，某科技企业为了训练 AI 筛查疾病的能力，首先需要对医疗影像数据进行处理，对病理切片进行分类和标注，以画框或描点的方式，将不同区域区别开来，并标注不同颜色以区分等级，为 AI 训练提供大量数据燃料。通过这种方式，该企业以深度学习预测前列腺癌的分类准确率已经达到 99.38%。

1.2.4　家居行业

智能家居在全球范围内呈现出强劲的发展势头，不仅基于日渐丰富的家居场景和日趋成熟的物联网技术，同时也离不开向前推进的图像识别、自然语言处理等技术。在助力智能家居发展中，数据标注主要应用矩形框标记人脸，进行人脸精细分割；对家居物品进行画框标记；通过描点的方式进行区域划分；采集语音并进行标注处理等。

在智能家居应用中，对于训练更“懂”人类的智能对话机器人，需要大量语料库支持训练，比如康奈尔电影对话语料库、Ubuntu 语料库和微软的社交媒体对话语料库[8]等都是比较常见的数据集，通过对以上数据进行标注处理，即可逐渐提升机器人回复的智能程度。

1.2.5　安防行业

目前，智能安防发展如火如荼。为了进一步提升安防应用的适用性，提高数据处理的速度与效率，推动安防从被动防御向主动预警发展，对数据标注的需求与日俱增。其中，人脸标注、视频分割、语音采集、行人标注等都是重要的数据标注应用。

在智能安防不断推进的过程中，生物识别技术已经越来越成熟，在日常监控、出入境管理、刑事案件侦查中都有着广泛应用。其中，对于数据

标注人员而言，需要做的正是对训练图片中人物的性别、年龄、肤色、表情、头发以及是否戴帽戴眼镜[9]等进行分类标注，或者对行人做标框处理，帮助机器获取快速识别能力。目前，天网系统应用动态人脸识别技术，使 1∶1 识别准确率达到 99.8%以上，同时可实现每秒比对 30 亿次，1 秒就能将全国人口“筛”一遍，2 秒便能将世界人口“筛”一遍[10]。

1.2.6 公共服务

对各种服务数据进行人工智能处理有助于提高公共服务水平与效率。以安防领域为例，目前大街小巷密布摄像头，但主要以记录与存储为主，大多用于事后侦查。随着数据标注的普遍应用，不断累积的海量标注数据，可以广泛用于人工智能训练，大大增强 AI+安防的合力，并可通过不断精进的人脸识别技术与视频行为分析技术，对监控画面进行实时分析，做到及时预警和响应。

具体而言，在海量标注数据的训练之下，智能安防可以识别人脸、分析表情、辨别身份，对公共场合人员进行快速统计。同时由于对特定行为进行了标注和训练，一旦监控视频中出现危险行为，系统将实时反馈和应对，避免潜在危险和损失。对于可疑人员，也可以加速侦查过程，更好地保障公共安全。

1.2.7 电子商务

在电商行业，数据标注能够帮助进一步深度挖掘数据集，建立客户全生命周期数据，预测需求趋势，优化价格与库存，最终达到精准营销的目的。通过互联网搜索指定内容答案的搜索完善、通过句子语境判断感情色彩的情绪分析以及人脸标注、语音采集等均为重要的数据标注应用。

对于电商数据而言，如虎鱼网络等专业系统，通过对产品打上结构化标签，包括品牌、颜色、型号、价格、款式、浏览量、购买量、用户评价等，建立 360° 的全景画像，从而为个性化推荐提供先决条件[11]。同时，该系统也可用于包括人口属性、购物偏好、消费能力、上网特征等在内的用户标签化处理，进一步建立用户兴趣图谱与用户画像，并通过智能推荐系统，推荐高转化的用户场景。

1.3 新职业–人工智能训练师

1.3.1 有多少智能，就有多少人工

人工智能一般由“数据”“算法”“应用”来支撑。对于机器学习而言，往往基于某个应用场景（比如人工智能程序 AlphaGo 主攻围棋），使机器通

过给定的数据学习参数总结规律、找出方向，进而提高算法（算法可理解为计算机解决问题的方法）。其中，数据成为当仁不让的关键点，输入数据，就会得到与该数据相对应的结果。

与此同时，机器学习又有监督学习与无监督学习之别。有监督学习首先通过训练样本找出规律，对模型进行优化，使其具有判断与预知能力，这是向“样本”学习的过程，其核心在于“分类”，多用于实际产品应用；而无监督学习缺少训练样本，直接通过数据进行建模分析，其核心在于“聚类”，主要用于探索研究。

换言之，只有在有监督学习的情况下，带有“标签”的数据才能成为模型优化的“老师”，也正是因为有监督学习，才需要大量经过标注的数据作为先验经验。然而，无论是数据标注，还是此前的数据采集、数据清洗与处理等，大多由人工完成，而数据处理的量级与质量又直接关系到机器的智能程度，也就是我们所说的“有多少智能，就有多少人工”。

举个例子，如果现在我们训练一个能够自动识别辣椒的人工智能程序，那么首先需要对大量含有辣椒的图片进行标注，确认是否带梗、颜色红绿等信息，将标注处理后的训练样本“喂”给等待训练的机器，授之以“渔”，使其基于算法框架自主学习，通过训练集学习，以测试集进行纠错，不断降低错误率，最终学成出师。在这个过程中，输入的数据样本越精确，规模越大，其处理效率与运作效率也越高。

1.3.2　让 AI 更懂人类的新职业

随着人工智能技术向各行各业纵深发展，致力让 AI 更懂人类的数据标注行业发展迅速，在短短几年内即被纳入国家职业分类目录，成为数字经济时代炙手可热的新职业。

2021 年，国家人力资源和社会保障部发布国家职业技能标准，进一步明确人工智能训练师（职业代码：4-04-05-05）包括数据标注员和人工智能算法测试员两个工种，并从下到上划分为五级-四级-三级-二级-一级共五个等级，分别对应初级工、中级工、高级工、技师以及高级技师，如图 1-8 所示。

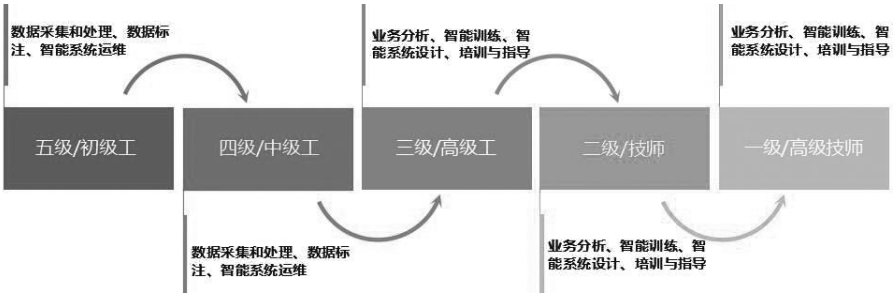


图 1-8　人工智能训练师五个职业等级

对于人工智能训练师 L5—L1 这五个等级，国家职业技能标准给出了具体的描述和要求，包括对应技能要求和相关知识要求，如表 1-1 至表 1-5 所示，为从业者提供了清晰的职业路径规划和指引。

表 1-1 五级/初级工[12]

职业功能	工作内容	技 能 要 求	相关知识要求
1. 数据采集和处理	1.1 业务数据采集	1.1.1 能够利用设备、工具等完成原始业务数据采集 1.1.2 能够完成数据库内业务数据采集	1.1.1 业务背景知识 1.1.2 数据采集工具使用知识 1.1.3 数据库数据采集方法
	1.2 业务数据处理	1.2.1 能够根据数据处理要求完成业务数据整理归类 1.2.2 能够根据数据处理要求完成业务数据汇总	1.2.1 数据整理规范和方法 1.2.2 数据汇总规范和方法
2. 数据标注	2.1 原始数据清洗与标注	2.1.1 能够根据标注规范和要求，完成对文本、视觉、语音数据清洗 2.1.2 能够根据标注规范和要求，完成文本、视觉、语音数据标注	2.1.1 数据清洗工具使用知识 2.1.2 数据标注工具使用知识
	2.2 标注后数据分类与统计	2.2.1 能够利用分类工具对标注后数据进行分类 2.2.2 能够利用统计工具，对标注后数据进行统计	2.2.1 数据分类工具使用知识 2.2.2 数据统计工具使用知识
3. 智能系统运维	3.1 智能系统基础操作	3.1.1 能够进行智能系统开启 3.1.2 能够简单使用智能系统	3.1.1 智能系统基础知识 3.1.2 智能系统使用知识
	3.2 智能系统维护	3.2.1 能够记录智能系统功能应用情况 3.2.2 能够记录智能系统应用数据情况	智能系统维护知识

表 1-2 四级/中级工[12]

职业功能	工作内容	技 能 要 求	相关知识要求
1.数据采集和处理	1.1 业务数据质量检测	1.1.1 能够对预处理后业务数据进行审核 1.1.2 能够结合人工智能技术要求，梳理业务数据采集规范 1.1.3 能够结合人工智能技术要求，梳理业务数据处理规范	1.1.1 业务数据质量要求和标准 1.1.2 业务数据采集规范和方法 1.1.3 业务数据处理规范和方法
	1.2 数据处理方法优化	1.2.1 能够对业务数据采集流程提出优化建议 1.2.2 能够对业务数据处理流程提出优化建议	1.2.1 数据采集知识 1.2.2 数据处理知识

续表

职业功能	工作内容	技 能 要 求	相关知识要求
2.数据标注	2.1 数据的归类和定义	2.1.1 能够运用工具，对杂乱数据进行分析，输出内在关联及特征 2.1.2 能够根据数据内在关联和特征进行数据归类 2.1.3 能够根据数据内在关联和特征进行数据定义	2.1.1 数据聚类工具知识 2.1.2 数据归纳方法 2.1.3 数据定义知识
	2.2 标注数据审核	2.2.1 能够完成对标注数据准确性和完整性审核，输出审核报告 2.2.2 能够对审核过程中发现的错误进行纠正 2.2.3 能够根据审核结果完成数据筛选	2.2.1 数据审核标准和方法 2.2.2 数据审核工具使用知识
3.智能系统运维	3.1 智能系统维护	3.1.1 能够维护智能系统所需知识 3.1.2 能够维护智能系统所需数据 3.1.3 能够为单一智能产品找到合适应用场景	3.1.1 知识整理方法 3.1.2 数据整理方法 3.1.3 智能应用方法
	3.2 智能系统优化	3.2.1 能够利用分析工具进行数据分析，输出分析报告 3.2.2 能够根据数据分析结论对智能产品的单一功能提出优化需求	3.2.1 数据拆解基础方法 3.2.2 数据分析基础方法 3.2.3 数据分析工具使用方法

表 1-3　三级/高级工[12]

职业功能	工作内容	技 能 要 求	相关知识要求
1.业务分析	1.1 业务流程设计	1.1.1 能够结合人工智能技术要求和业务特征，设计整套业务数据采集流程 1.1.2 能够结合人工智能技术要求和业务特征，设计整套业务数据处理流程 1.1.3 能够结合人工智能技术要求和业务特征，设计整套业务数据审核流程	1.1.1 业务数据相关流程设计工具知识 1.1.2 业务数据相关流程设计知识
	1.2 业务模块效果优化	1.2.1 能够结合业务知识，识别业务流程中单一模块的问题 1.2.2 能够结合人工智能技术设计业务模块优化方案并推动实现	1.2.1 业务分析方法 1.2.2 业务优化方法
2.智能训练	2.1 数据处理规范制定	2.1.1 能够结合人工智能技术要求和业务特征，设计数据清洗和标注流程 2.1.2 能够结合人工智能技术要求和业务特征，制定数据清洗和标注规范	2.1.1 智能训练数据处理工具原理和应用方法 2.1.2 智能训练数据处理知识

续表

职业功能	工作内容	技 能 要 求	相关知识要求
2.智能训练	2.2 算法测试	2.2.1 能够维护日常训练集与测试集 2.2.2 能够使用测试工具对人工智能产品的使用进行测试 2.2.3 能够对测试结果进行分析，编写测试报告 2.2.4 能够运用工具，分析算法中错误案例产生的原因并进行纠正	2.2.1 人工智能测试工具使用方法 2.2.2 算法训练工具基础原理和应用方法
3.智能系统设计	3.1 智能系统监控和优化	3.1.1 能够对单一智能产品使用的数据进行全面分析，输出分析报告 3.1.2 能够对单一智能产品提出优化需求 3.1.3 能够为单一智能产品的应用设计智能解决方案	3.1.1 数据拆解高阶方法 3.1.2 数据分析高阶方法 3.1.3 单一产品智能解决方案设计方法
	3.2 人机交互流程设计	3.2.1 能够通过数据分析，找到单一场景下人工和智能交互的最优方式 3.2.2 能够通过数据分析，设计单一场景下人工和智能交互的最优流程	3.2.1 人机交互流程设计知识 3.2.2 人机交互流程设计工具相关知识
4.培训与指导	4.1 培训	4.1.1 能够编写初级培训讲义 4.1.2 能够对五级/初级工、四级/中级工开展知识和技术培训	4.1.1 培训讲义编写知识 4.1.2 培训教学知识
	4.2 指导	4.2.1 能够指导五级/初级工、四级/中级工解决数据采集、处理问题 4.2.2 能够指导五级/初级工、四级/中级工解决数据标注问题	4.2.1 实践教学方法 4.2.2 技术指导方法

表 1-4 二级/技师[12]

职业功能	工作内容	技 能 要 求	相关知识要求
1.业务分析	1.1 业务框架与流程设计	1.1.1 能够综合业务流程及重难点，结合人工智能技术构建合理的业务框架 1.1.2 能够综合业务流程及重难点，结合人工智能技术构建合理的业务流程	1.1.1 业务流程构建工具原理和应用方法 1.1.2 业务流程构建知识
	1.2 业务场景挖掘	1.2.1 能够在业务中挖掘智能系统应用的潜在机会点及隐藏价值 1.2.2 能够结合人工智能技术对新业务场景提出解决方法	1.2.1 数据分析方法 1.2.2 数据运营方法
2.智能训练	2.1 算法测试	2.1.1 能够结合业务特征，构建算法的高质量训练集，并成为算法的核心竞争力 2.1.2 能够结合业务特征，构建算法的黄金测试集，并作为算法上线前的质量保障 2.1.3 能够结合业务特性，设计合理的测试方案	2.1.1 人工智能算法基础知识 2.1.2 算法测试工具原理和应用方法

续表

职业功能	工作内容	技 能 要 求	相关知识要求
2.智能训练	2.2 智能训练流程优化	2.2.1 能够根据日常算法模型的训练，提出训练产品优化需求并推动实现 2.2.2 能够根据日常算法模型的训练，提出训练方法的新思路	2.2.1 算法训练工具设计和优化方法 2.2.2 算法训练方法优化方法
3.智能系统设计	3.1 智能产品应用解决方案设计	3.1.1 能够在某一业务领域中设计包含多个智能产品的解决方案并推动实现 3.1.2 能够基于某一业务领域情况，结合多个智能产品设计新的全链路智能应用流程	3.1.1 业务领域知识 3.1.2 多智能产品解决方案设计方法
	3.2 产品功能设计以及实现	3.2.1 能够将解决方案转化成产品功能需求 3.2.2 能够推动产品功能需求实现并达成项目目标	3.2.1 产品需求梳理方法 3.2.2 项目管理知识
4.培训与指导	4.1 培训	4.1.1 能够编写培训计划 4.1.2 能够对三级/高级工及以下级别人员开展知识和技术培训	4.1.1 培训计划编制知识 4.1.2 进阶培训教学知识
	4.2 指导	4.2.1 能够制定业务指导方案 4.2.2 能够对三级/高级工及以下级别人员培训学习效果进行评估	4.2.1 业务指导方案编制方法 4.2.2 效果评估方法

表 1-5　一级/高级技师[12]

职业功能	工作内容	技 能 要 求	相关知识要求
1.业务分析	1.1 业务设计	1.1.1 能够根据复杂业务场景和跨业务单元场景的深入理解，搭建业务分析框架 1.1.2 能够结合人工智能技术为所负责的业务线提出具有前瞻性的业务发展规划建议	1.1.1 业务指标定义知识 1.1.2 业务指标的管理方法 1.1.3 业务发展规划设计方法
	1.2 业务创新	1.2.1 能够利用人工智能技术，对现有业务流程重构，提高业务在行业领域竞争力 1.2.2 能够结合先进的人工智能技术，在业务流程中发现创新点并整合，推动行业领域的创新 1.2.3 能够结合人工智能技术，前瞻性的洞察行业业务战略方案	1.2.1 人工智能技术相关知识 1.2.2 流程设计创新方法
2.智能训练	2.1 算法测试	2.1.1 能够根据对算法的前瞻性，制定智能训练的整体产品能力矩阵 2.1.2 能够根据对算法的前瞻性，制定训练平台的整体迭代优化方案 2.1.3 能够制定训练集以及测试集的标准	2.1.1 智能训练工具高阶原理和应用方法 2.1.2 智能训练技巧和方法 2.1.3 人工智能算法高阶知识

续表

职业功能	工作内容	技 能 要 求	相关知识要求
2.智能训练	2.2 智能训练流程优化与产品化	2.2.1 能够对复杂的智能系统进行完整的测试和训练，并做出报告编写 2.2.2 能够结合人工智能技术，对智能训练的完整体系提出新思路，新方向，并推动产品更新	2.2.1 人工智能技术创新方法 2.2.2 智能训练产品原理和方案优化设计方法
3.智能系统设计	3.1 智能产品应用解决方案设计	3.1.1 能够在复杂业务领域中设计包含多个智能产品的解决方案并推动实施 3.1.2 能够跨多业务领域设计智能产品应用方案，解决业务问题	3.1.1 智能行业和业务知识 3.1.2 跨多业务领域智能解决方案设计方法
	3.2 平台化推广	3.2.1 能够将方法论沉淀，应用到智能算法或者知识体系中，并给行业带来变革 3.2.2 能够独立统筹并推动项目进行，推动多个智能产品的一系列运营，实现项目目标	3.2.1 项目管理方法 3.2.2 产品运营方法
4.培训与指导	4.1 培训	4.1.1 能够制定培训体系规划 4.1.2 能够对二级/技师及以下级别人员开展管理方法培训	4.1.1 培训体系构建方法 4.1.2 管理培训知识
	4.2 指导	4.2.1 能够制定业务指导策略体系 4.2.2 能够对二级/技师进行业务指导	4.2.1 业务指导策略体系编制方法 4.2.2 人工智能训练前沿理论知识

1.3.3 最后一批人工智能的“老师”

有多少智能，就有多少人工，在一定意义上，人工智能工程师可以看作是人工智能的老师，因为他们标注的各种图像、文本与语音教会了机器学习，且标注的数量和质量与机器学习成果直接关联。按照这一思路推演，如果人工智能需要学习新本事，需要不断提升和完善，那么这一职业就将伴随其存在下去。

然而，随着人工智能的疯狂生长，训练好的 AI 模型反哺人工标注，当数据足够丰富和完善的时候，只需将数据库之外的图片、语音和文本等，交给人工智能进行识别，并基于数据库的识别数据调整参数和验证，就可以达到更高的精准度。

与此同时，人工智能将逐渐实现由弱人工智能向强人工智能直至超人工智能的转变，大量人类岗位将由机器人替代，青出于蓝而胜于蓝，最终“学生”将全面超越“老师”。在智能升级的过程中，随着有监督学习向无

监督学习或迁移学习的转变，数据标注的需求也将大幅度削减，即人工标注最终可能将不复存在。

不过，目前无监督学习等只是处于探索阶段的新算法，并没有大规模的商业落地。为此，即使最终将退出历史舞台，人工智能训练师也是陪伴人工智能成长壮大的最后一批“老师”。

除人工标注之外，AI 辅助工具也逐渐应用到具体的标注过程中，比如谷歌推出的“流体标注”工具。通常而言，在 COCO+Stuff 数据集中，标记一个图像需要 19 分钟，而标记整个数据集需要 53000 小时[13]，而在谷歌对“流体标注”的展示中，在机器辅助之下，“流体标注”能够清晰圈出目标轮廓和背景，完成数据标注过程。

如图 1-9 所示，图片中列展示的是在 COCO 数据集中对 3 张图片的传统手动标记，而右列则是通过“流体标注”对图片进行的标记。不难看出，“流体标注”与手动标记的呈现效果基本上相差无几，除了智能程度得到大幅度提升之外，标注数据集的生成速度也得以显著提高，可以达到原来的 3 倍。

图 1-9　手动标记和流体标注对比[13]

1.4　数据越多，智能越好

谷歌和美国卡内基梅隆大学联合发布的一篇论文中明确指出，深度学习的成功归功于：① 高容量的模型；② 越来越强的计算能力；③ 可用的大规模标签数据[13]。然而在此前的研究中发现，2012—2016 年计算力（得

益于 GPU）与模型尺寸不断增长，但每年数据集规模却基本保持不变，如图 1-10 所示。

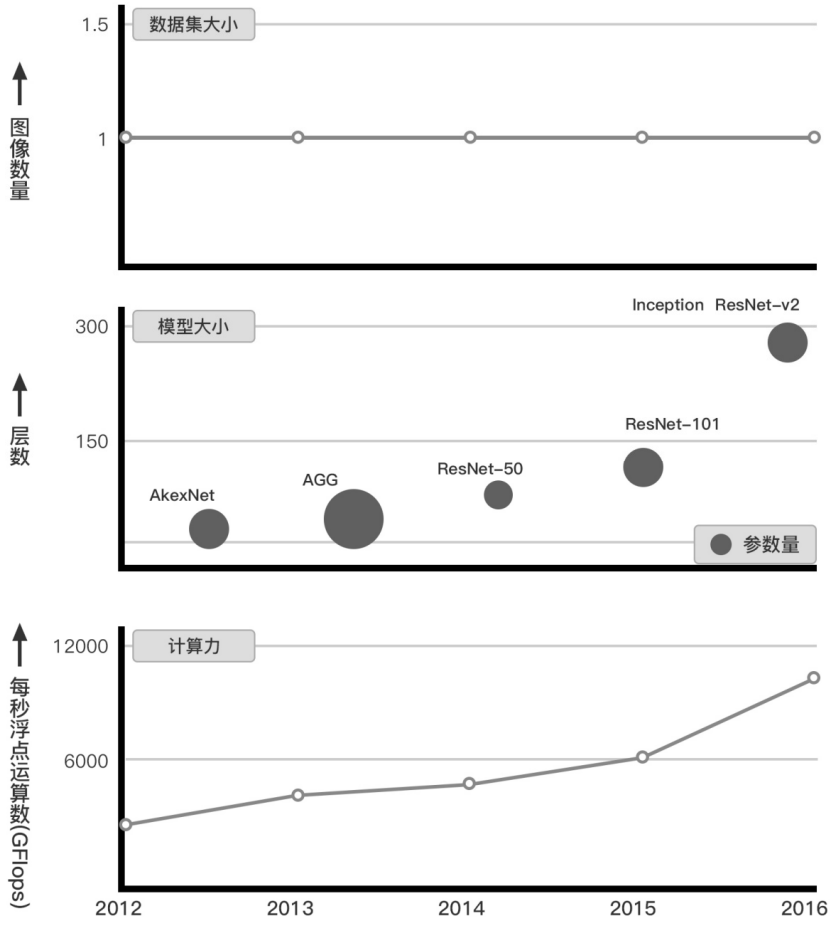


图 1-10　模型尺寸、计算力与数据规模对比[14]

这时研究人员提出猜想，当数据规模成百倍成千倍增长时，人工智能研究的精度与准确性会怎么改变呢？是存在一定的“天花板”，还是精度与准确度会随着数据量的增长越来越高？为了得到确实的结果，研究人员应用谷歌建立的内部数据集——JFT-300M（数据是 ImageNet 的 300 倍，含有超过 10 亿个标签）进行研究。

通过最终实验结果发现，任务性能与训练数据之间关系紧密，大规模数据有助于表征学习，同时随着训练数据的数量级增长，模型性能呈线性增长，大规模的数据集对于预训练而言大有助益，如图 1-11 所示。

不难看出，欣欣向荣的人工智能行业直接拉动了数据标注行业的崛起和发展。随着感知智能向认知智能的转变，对于标注数据的维度与细化程度也提出了更高要求。与此同时，在有监督学习之下，海量高准确率的标

注数据进一步推动了人工智能的行业落地，标注的数据越多，智能水平也越高。

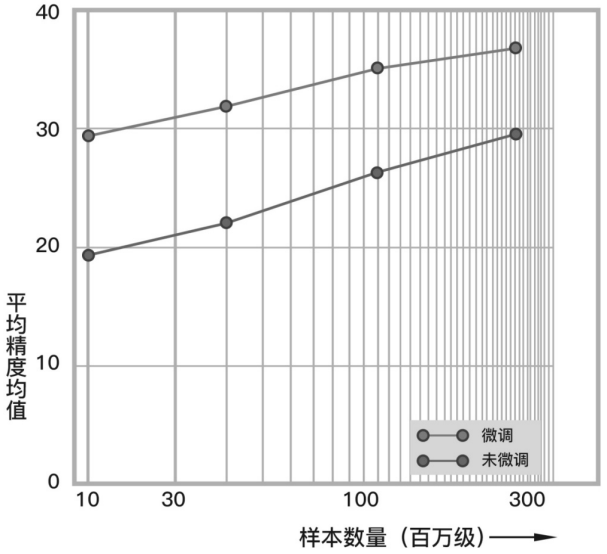


图 1-11　测试性能随数据量呈线性增长[14]

1.5　作业与练习

1．如何理解数据标注与人工智能的关系？
2．什么是数据标注？
3．数据标注对象可以划分为哪几类？
4．数据标注流程包括哪些环节？
5．数据标注有哪些应用场景？
6．如何理解“有多少智能，就有多少人工”？
7．人工智能训练师包括哪几个职业等级？
8．数据量级与智能程度之间存在怎样的联系？

参考文献

[1] 环球网. 就业新观察丨数字经济催生更多新职业 到 2025 年带动就业人数将达 3.79 亿[DB/OL].（2022-03-31）[2022-05-07]. https://baijiahao.baidu.com/s?id=1728807346756307367&wfr=spider&for=pc.

[2] 钱塘数据. 电子标准院：人工智能标准化白皮书（2018 版）[DB/OL].（2018-10-26）[2022-05-10]. https://cloud.tencent.com/developer/article/1359067.

[3] 国务院. 国务院关于印发“十四五”数字经济发展规划的通知[EB/OL].（2022-01-12）[2022-05-12]. www.gov.cn/zhengce/zhengceku/2022-01/12/

content_5667817.htm.

[4] 锐观 CC．2023—2028 年中国数据标注产业全景调查及投资咨询报告[DB/OL].（2022-06-23）[2022-07-2]．https://view.inews.qq.com/a/20220623A030HA00.

[5] 阿里云．深入浅出看懂 AlphaGo Zero - PaperWeekly 第 51 期[DB/OL].（2017-10-24）[2022-07-22]．https://developer.aliyun.com/article/226363.

[6] 精灵标注．精灵标注助手[DB/OL]．[2022-07-23]．www.jinglingbiaozhu.com/?b_scene_zt=1.

[7] 中华人民共和国人力资源和社会保障部．人工智能训练师 国家职业技能标准（2021 年版）[S/OL]．[2022-07-25]．www.mohrss.gov.cn/wap/zc/zqyj/202106/W020210617509883457681.pdf.

[8] 利荣．Scale 推出传感器融合标注 API，为自动驾驶技术更快注入数据燃料[DB/OL].（2018-03-07）[2022-07-26]．https://www.leiphone.com/category/transportation/mlpbK1Q4vUrSzU80.html.

[9] AI 科技大本营．实战|让机器人替你聊天，还不被人看出破绽？来，手把手教你训练一个克隆版的你[DB/OL].（2017-08-23）[2022-07-27]．https://mp.weixin.qq.com/s/jNTVKGTgNlucEpPB9vZvkA.

[10] 赟尘．谈谈数据标注那些事[DB/OL].（2017-11-24）[2020-07-30]．http://www.woshipm.com/pd/856172.html.

[11] 中国江苏网．“天网”已应用全国 16 省市　人脸识别技术助力安防[DB/OL]．（2018-03-23）[2022-07-22]．https://baijiahao.baidu.com/s?id=1595690163111887808&wfr=spider&for=pc.

[12] 凤凰网科技．第一批被 AI 累死的人[DB/OL].（2018-07-15）[2022-08-01]．http://tech.ifeng.com/a/20180715/45063971_0.shtml.

[13] 新智元．谷歌推出“流体标注”AI 辅助工具，图像标注速度提升 3 倍！（附论文）[DB/OL].（2018-10-23）[2022-08-02]．http://www.sohu.com/a/270697508_473283.

[14] 【10 亿+数据集，ImageNet 千倍】深度学习未来，谷歌认数据为王[DB/OL].（2017-07-12）[2022-08-01]．http://www.sohu.com/a/156480210_473283.