第 3 章 开源数据的利用与整合

【教学目标】

（1）掌握开源数据的定义、优势（如降低成本、促进创新）和局限性（如数据偏差、更新不及时），理解其在 AI 项目中的价值；

（2）熟悉主流开源数据集（如 ImageNet、IMDb、LibriSpeech、UCF101、Common Crawl）的特点、规模和应用场景；

（3）理解常见数据许可协议（如 CC BY、CC BY-SA、CC0）的授权范围和合规要求，具备评估数据集合法性的能力；

（4）能够使用 Kaggle API 下载开源数据集并进行预处理（如格式转换、清洗），验证数据完整性（SHA-256）；

（5）设计开源数据融合方案，结合自有数据提升数据集规模和模型泛化能力，解决偏差和适配性问题。

在人工智能的发展过程中，开源数据扮演了至关重要的角色。它为研究人员和工程师提供了宝贵的资源，加速了模型的开发和创新。本章将详细探讨开源数据的概念、优势与案例及局限，介绍常用的开源数据集，以及如何有效地获取、使用和整合开源数据。

电子工业出版社有限公司 版权所有

3.1 开源数据的概念与意义

人工智能模型的性能在很大程度上取决于底层所使用的数据的质量和数量。随着 AI 技术的不断发展，数据驱动的模型训练、验证和评估已成为业界标准。在此背景下，“开源数据”这一概念日益受到关注，并在学术研究和产业实践中发挥着举足轻重的作用。

3.1.1 开源数据的概念

开源数据（Open Data）是指在合法许可协议下，免费或自由获取并可用于数据分析、模型训练和商业用途的数据。与传统的私有数据资源不同，开源数据通常以明确的授权协议（如 Creative Commons、Open Data Commons 或 Public Domain License）进行发布，使研究人员、开发者和企业无须额外付费或办理复杂授权手续即可获取和使用。这种开放共享的模式不仅降低了数据获取门槛，也为创新实践、研究合作和学术成果的验证与对比提供了有利条件。

开源数据具有以下特征：

- 免费和自由获取：用户无须付费即可下载和使用数据。

- 透明和可溯源：数据的来源、采集方式和处理流程通常有清晰的文档或说明，便于验证和追溯。
- 可再分发和再利用：用户在遵守相应许可协议的前提下，可对数据进行再加工、再分发、衍生作品创建和商用化应用。

3.1.2 开源数据的优势与案例

开源数据具有规模大、来源广泛的特点。例如，著名的 ImageNet 数据集推动了计算机视觉领域的飞速发展，GPT 系列模型利用大量开源文本数据成为当前 NLP 领域的代表性成果。

1．开源数据的优势

（1）降低数据获取成本。

与自建数据采集管线、付费购买数据相比，开源数据能够有效降低成本，为中小企业、初创公司、独立研究者和学术团队创造更公平的起点。例如，创业公司在启动阶段往往预算有限，通过使用成熟的开源数据集，可以快速验证模型原型而无须投入大量资源。

（2）促进研究创新与社区合作。

开源数据为学术界和产业界搭建了知识共享平台。全球研究人员可以基于同一批数据开展不同方向的研究，从而加速学术成果的迭代与演进。这种共享机制也有利于模型结果的可比性与可重复性，研究人员能够在相同数据集上对比不同算法的优劣，进而推动技术进步和方法学创新。

（3）标准化基准测试与性能评估。

标准的开源数据集为 AI 模型性能评估提供了统一的比较基准。研究人员和工程师可在相同的数据集上测试模型，从而客观评估算法改进或新方法的有效性。这种共同基准有助于建立领域共识，并提高行业技术交流与协同发展的效率。

（4）缩短实验周期与开发周期。

使用成熟的开源数据集可避免从零开始的数据生产，使项目团队将更多精力集中在模型开发、特征提取和算法优化上，从而加快产品研发和迭代的速度。

2．典型案例：ImageNet 数据集

作为计算机视觉领域最著名的开源数据集之一，ImageNet 包含超过 1400 万张标注精确的图像，并以层次化的类别结构（基于 WordNet）组织数据。这一庞大的开源数据集为图像分类、目标识别、特征学习等研究方向提供了标准化的基准。其显著特征包括：

- 创新促进：ImageNet 大规模视觉识别挑战赛（ImageNet Large Scale Visual Recognition Challenge，ILSVRC）让全球研究人员和公司在同一数据集上展示模型性能，从而促进了深度卷积神经网络（CNN）的崛起和快速进步。
- 降低门槛：无论是初学者还是资深研究人员，都可轻松获取 ImageNet 数据，在同一平台下快速实现模型训练与性能验证。
- 产业应用：许多科技公司在 ImageNet 预训练模型的基础上进行迁移学习，为各种垂直领域应用（如医疗影像分析、无人驾驶视觉系统）提供了坚实的数据基础与性能参考。

电子工业出版社有限公司 版权所有

3.1.3 开源数据的局限

尽管开源数据具有诸多优势，但在实际应用中仍存在一定局限，需要在使用过程中保持谨慎和批判性思维。开源数据的具体局限体现在以下几个方面：

- 数据质量与合规性不均衡：开源数据的生产过程可能缺乏严谨的数据工程化标准。部分数据集存在质量参差不齐、标注不一致、缺失元数据、重复数据和错误标注的问题。对于商业应用场景，用户可能需要投入额外的精力对开源数据进行清洗、标注质量检查和后期加工。
- 数据偏差与不公平性：开源数据往往由特定的组织或研究团队在特定地域、文化背景下收集，可能在数据分布、样本多样性和标注标准上存在偏差。例如，ImageNet 等开源数据集中仍可能存在类别不均衡、区域与文化偏向等问题。这会导致模型在多元化应用场景中表现不稳定，甚至引发不公平决策与伦理争议。
- 更新与维护不及时：某些开源数据集在发布后缺乏持续维护，无法反映最新的现实世界动态。例如，一个用于城市交通预测的开源数据集若长期不更新，将难以反映最新的路网变化、出行模式转变。缺乏及时更新的开源数据可能降低模型的实际应用价值。
- 特定场景适配性不足：通用的开源数据集往往覆盖广泛，但对某一特定细分领域的针对性不强。对高度专业化的场景（如医疗影像分析、工业缺陷检测）而言，开源数据集可能缺乏足够的行业特征与数据细节，需要进行额外的数据扩充或与私有数据融合才能满足特定业务需求。

例如，虽然 ImageNet 在图像分类任务中发挥了巨大作用，但在更细粒度的任务上（如医疗影像的特定病灶识别、工业检测中微小缺陷的发现），ImageNet 的通用性可能存在不足。研究者还需要另外收集细分领域的数据集，或对 ImageNet 预训练模型进行大量的微调和数据增强，才能取得满意的性能。

开源数据是 AI 时代知识与资源共享生态中的重要一环，为研究者与从业者提供了低成本、高透明度的共享资源，以及标准化的对比基准和快速迭代的土壤。以 ImageNet 为代表的广泛应用的开源数据集，不仅推动了学界与产业界的技术演进，也为初创企业和独立研究者打开了大门。然而，开源数据并非完美无缺，其数据质量、偏差、维护更新和特定场景适配性等方面的局限性仍需在实践中予以警惕和改进。通过合理地选择、清洗与整合开源数据，并结合私有数据和定制数据工程流程，才能最大限度地发挥开源数据在人工智能数据工程中的潜在价值。

3.2 常用的开源数据集

本节将介绍几种常见的开源数据集。在选择开源数据集时应考虑数据质量、数据规模、许可协议和更新频率，以满足不同场景的需求。

电子工业出版社有限公司 版权所有

3.2.1 图像数据集

1．MNIST：手写数字识别

MNIST 数据集包含 6 万张训练图像和 1 万张测试图像，每张图像都是一个 28×28 像素的手写数字黑白图像，标签为数字 0～9；广泛用于图像分类的入门实验，适合初学者用于机器学习和深度学习模型的训练与验证。

2．CIFAR-10/100：物体分类

（1）CIFAR-10。

CIFAR-10 包含 6 万张 32×32 像素的彩色图像，分为 10 个类别，每个类别有 6000 张图像；用于物体识别和图像分类任务的研究及模型评估。

（2）CIFAR-100。

与 CIFAR-10 类似，但 CIFAR-100 包含 100 个类别，每个类别有 600 张图像；用于更细粒度的分类任务，挑战模型的分类能力。

3．ImageNet：大规模图像分类

ImageNet 是一个大规模的图像数据集，包含超过 1400 万张带标注的图像，涵盖 20000 个类别（核心子集分为 1000 个类别）；用于训练深度卷积神经网络，推动了计算机视觉的发展；ImageNet 大规模视觉识别挑战赛（ILSVRC）是计算机视觉领域的重要赛事，促进了模型的创新和性能提升。

4．案例分析：ImageNet 数据集对深度学习的推动作用

2012 年，AlexNet 模型在 ILSVRC 中取得了突破性的成绩，显著优于传统方法，对计算机视觉产生重大影响：

- 推动深度学习发展：AlexNet 的成功证明了深度卷积神经网络在图像识别中的强大能力，引发了学术界和工业界的广泛关注。
- 模型创新：随后出现了 VGG、GoogLeNet、ResNet 等更深更复杂的模型，不断刷新图像识别的性能。
- 应用扩展：深度学习的方法被应用于目标检测、图像分割、风格迁移等多个领域。

3.2.2 文本数据集

1．IMDb：电影评论情感分析

IMDb（Internet Movie Database）包含 5 万条经过标注的电影评论文本，分为正面和负面两类，平均每条评论 200 多个词；用于文本分类、情感分析等自然语言处理任务。

2．SST-2：句子情感分类

SST-2（Stanford Sentiment Treebank v2）是一个对单句子进行情感分类的数据集，标签分为正面和负面；用于细粒度的情感分析，评估模型对短文本情感的识别能力。

3．20 Newsgroups：文本分类

20 Newsgroups 包含约 2 万篇新闻组的帖子，分为 20 个不同的主题类别；用于文本分类、主题建模等任务。

电子工业出版社有限公司 版权所有

4．案例分析

使用 IMDb 数据集训练情感分析模型的实战，包含了典型的模型训练和评估过程。

（1）数据加载。

使用 Python 的 TorchText 或 Keras 等库加载 IMDb 数据集。

（2）数据预处理。

- 分词：将文本划分为单词或词组。
- 去除停用词：删除无意义的词汇，如“the”“and”等。
- 词向量表示：将文本转换为数字表示，可以使用词袋模型、TF-IDF 或预训练的词向量。

（3）模型构建。

可以使用朴素贝叶斯、支持向量机、LSTM、BERT 等模型。

（4）模型训练。

- 设置损失函数和优化器。
- 训练模型，在验证集上评估性能。

（5）模型评估。

- 准确率、精确率、召回率、F1 值等指标。
- 混淆矩阵：分析模型的分类错误情况。

（6）结果。

- 模型性能：在测试集上达到较高的准确率，证明了模型的有效性。
- 经验总结：数据预处理和模型选择对情感分析的效果有显著影响。

3.2.3　音频数据集

音频数据集通常用于语音识别、声纹识别等场景。例如，LibriSpeech 常用于语音识别模型训练，UrbanSound8K 用于环境声音分类任务。

1．LibriSpeech：语音识别

LibriSpeech 是一个大型的英语语音数据集，包含约 1000 小时的读书语音，文本转录准确；用于训练和评估自动语音识别（ASR）系统。

2．VoxCeleb：说话人识别

VoxCeleb 包含来自 YouTube 上名人的语音片段，包含超过 6000 名说话者；用于说话人识别、验证和分离任务。

3．案例分析

以下是 LibriSpeech 在语音识别系统中的应用案例，包含了典型的模型训练和评估过程。

（1）数据准备。

- 下载数据集：从官方网站获取 LibriSpeech 数据集。
- 数据格式：音频文件为 WAV 格式，文本为对应的转录文件。

（2）数据预处理。

- 特征提取：从音频中提取 MFCC（梅尔频率倒谱系数）或其他声学特征。
- 数据增强：加入噪声、变速等方式，增加数据多样性。

电子工业出版社有限公司 版权所有

（3）模型构建。

传统的 HMM-GMM 模型，或基于深度学习的 CTC（连接时序分类）模型、Transformer。

（4）模型训练。

- 设置损失函数和优化器。
- 训练模型，监控损失和错误率。

（5）模型评估。

- 词错误率（WER）：评估识别结果与真实转录的差异。
- 可视化结果：分析模型在不同语音上的表现。

（6）结果。

- 模型性能提升：利用大型高质量的 LibriSpeech 数据集，模型在语音识别任务上取得了较好的效果。
- 启示：高质量的数据对语音识别系统的性能至关重要。

3.2.4 视频数据集

1．UCF101：动作识别

UCF101 包含约 13320 个视频剪辑，分为 101 个动作类别，如跳舞、骑自行车等；用于视频动作识别、视频分类等任务。

2．Kinetics：大型视频动作数据集

Kinetics 由 DeepMind 提供，包含超过 65 万个视频片段，分为 400 个动作类别；适用于深度学习模型的训练和评估，推动视频理解领域的发展。

3．案例分析

以下是利用UCF101 训练动作识别模型的效果评估案例，包含了典型的模型训练和评估过程。

（1）数据准备。

下载 UCF101 数据集，并解压视频文件。

（2）数据预处理。

- 帧提取：从视频中提取关键帧，或者将视频转换为帧序列。
- 数据增强：随机剪辑、水平翻转、尺度变换等。

（3）模型构建。

可选择 3D 卷积神经网络（如 C3D）、双流网络、时空注意力模型等。

（4）模型训练。

- 设置损失函数和优化器。
- 训练模型，在验证集上评估性能。

（5）模型评估。

- Top-1 和 Top-5 准确率：评估模型的分类效果。
- 混淆矩阵：分析模型对不同动作类别的识别情况。

（6）结果。

- 模型取得良好性能：在 UCF101 数据集上达到了较高的准确率。
- 挑战与改进：部分动作类别容易混淆，需要进一步改进模型的时空特征提取能力。

电子工业出版社有限公司 版权所有

3.2.5 网页数据集

1．Common Crawl：大规模网页数据

Common Crawl 是一个包含数十亿网页的开放数据集，提供网页的原始 HTML、文本提取等；每月更新一次，单次抓取数据达几十 TB，通常需要利用分布式处理平台（如 Hadoop、Spark）或云平台进行数据的抽取、过滤和分析，以提高处理效率；用于训练大型的语言模型、信息检索、网页分析等。

2．Wikipedia Dump：维基百科文本

Wikipedia Dump 是维基百科的完整数据，包括所有的条目和修订历史；英文维基百科压缩文件约为 25GB，中文维基百科压缩文本约为 2GB，可从维基媒体官网获取，适合中文自然语言处理研究；用于知识图谱构建、文本生成、问答系统等。

3．案例分析

大型预训练语言模型（如 GPT 系列）需要海量的文本数据进行训练。以下是使用 Common Crawl 数据集训练预训练语言模型的案例，包含了典型的模型训练和评估过程。

（1）数据准备。

从 Common Crawl 官方网站获取压缩的网页数据。

（2）数据预处理。

- 文本提取：从 HTML 中提取纯文本内容，去除标签和脚本。
- 数据清洗：过滤广告、重复内容和低质量文本。

（3）模型训练。

- 选择模型：Transformer 架构的语言模型。
- 训练策略：使用分布式训练，加速模型的训练过程。

（4）模型评估。

- 下游任务测试：在文本分类、问答、摘要等任务上评估模型的效果。
- 困惑度（Perplexity）：衡量模型对语言的预测能力。

（5）结果。

- 模型具备强大的语言理解和生成能力，在多个任务上表现出色。
- 贡献：促进了自然语言处理领域的进步，为下游应用提供了有力的支持。

3.3 开源数据的获取与使用

3.3.1 数据许可协议解读

数据许可协议规定了数据的使用权限和限制，例如 Creative Commons（CC）许可协议，包括 CC BY、CC BY-SA 等，明确了数据的使用和再分发方式。在使用开源数据时，必须遵守数据的许可协议，确保合法合规。

电子工业出版社有限公司 版权所有

1．常见协议

（1）CC BY（署名）。

- 要求：使用者必须署名，注明原作者和来源。
- 限制：可以自由使用、修改和分发，包括商业用途。

（2）CC BY-SA（署名-相同方式共享）。

- 要求：在 CC BY 的基础上，要求衍生作品也采用相同的协议。
- 限制：保持开源精神，防止闭源化。

（3）CC0（公共领域）。

- 要求：作者放弃所有版权，数据进入公共领域。
- 限制：无，任何人都可以自由使用。

2．使用限制和合规要求

- 禁止侵权：不得违反版权法，尊重原作者的权利。
- 遵守隐私法规：涉及个人数据时，需符合 GDPR、CCPA 等隐私保护框架。
- 避免不当使用：不得用于非法或不道德的目的，如歧视、欺诈等。

3．案例分析：违反数据许可协议的法律后果

某公司在未遵守 CC BY-SA 协议的情况下，将一个开源数据集用于商业产品，且未公开其修改后的数据和代码。

（1）法律问题。

- 违反许可协议：未遵守“相同方式共享”的要求，侵犯了原作者的权利。
- 侵权责任：原作者有权要求其停止侵权行为，赔偿损失。

（2）后果。

- 法律诉讼：公司被起诉，面临高额的赔偿和法律费用。
- 声誉受损：侵权行为被曝光后，公司声誉受到严重影响。

（3）启示。

- 重视许可协议：在使用开源数据时，必须仔细阅读并遵守相关许可协议的条款。
- 合规经营：确保所有的数据使用都符合法律法规，避免法律风险。

3.3.2 数据预处理与适配

数据预处理包括格式转换、清洗和标准化等操作，例如，将网页 HTML 格式数据转换为纯文本、去除噪声内容（如广告链接、脚本代码等），以便进行自然语言处理。

1．数据格式转换

（1）目的。

将开源数据集转换为项目所需的格式，方便后续处理和模型训练。

（2）方法。

- 文件格式转换：如从 CSV 转换为 JSON，或图像格式的转换（JPEG、PNG、BMP、TIFF 等）。
- 数据结构调整：重新组织数据的层次和字段，满足模型输入的要求。

电子工业出版社有限公司 版权所有

2．标准化处理

（1）统一数据规格，确保数据的尺寸、类型、单位等一致，例如：

- 图像：调整尺寸、归一化像素值。
- 文本：统一文字、字符的编码格式（如 UTF-8），规范化字符（小写化、去除特殊符号）等。

（2）处理缺失值和异常值，如填充或删除缺失数据，处理离群点。

3．实践活动

对开源数据集进行预处理，使其适用于特定任务，熟悉文本数据的预处理与模型训练过程。具体任务是使用 20 Newsgroups 数据集进行文本分类，主要目标是熟悉预处理数据的方法和过程，使其适用于训练神经网络模型。

（1）数据加载。

使用 sklearn.datasets 中的 fetch_20newsgroups 函数加载数据。

（2）数据预处理。

- 文本清洗：去除 HTML 标签、特殊字符和数字。
- 分词：使用 nltk 或 spaCy 等库进行分词。
- 去除停用词：删除常见但无意义的词汇。
- 词干提取或词形还原：规范化词语形式。

（3）特征表示。

- 词袋模型（Bag of Words）：将文本转换为词频向量。
- TF-IDF：计算词语的 TF-IDF 值，突出重要词汇。
- 词向量嵌入：使用预训练的词向量（如 GloVe、Word2Vec）。

（4）数据划分。

将数据集划分为训练集、验证集和测试集。

（5）模型训练。

选择适当的模型，如朴素贝叶斯、支持向量机、神经网络。

（6）结果。

- 成功训练模型，并在测试集上取得了良好的分类准确率。
- 经验总结：数据预处理对文本分类的效果有显著影响，需根据任务需求进行调整。

3.3.3　数据质量评估

数据质量评估包括准确性、一致性和完整性等内容，常用自动评估工具如 Great Expectations、TensorFlow Data Validation（TFDV）可自动检测数据中的异常值、缺失值等问题。

1．评估指标

（1）准确性。

- 定义：数据是否真实、正确地反映了客观事实。
- 评估方法：随机抽样检查，验证数据与原始来源是否一致。

（2）一致性。

- 定义：数据在不同部分之间是否保持一致，没有冲突。

电子工业出版社有限公司 版权所有

- 评估方法：检查数据格式、单位、命名等是否统一。

（3）完整性。

- 定义：数据是否完整，是否存在缺失或不完整的记录。
- 评估方法：统计缺失值的数量和比例，分析缺失模式。

2．实践活动

选择一个开源数据集，如 Kinetics，建立数据质量评估报告，提出改进方案。

（1）数据检查。

- 准确性：核对部分视频的标签，确保标注正确。
- 一致性：检查视频格式、分辨率、帧率是否统一。
- 完整性：统计缺失或损坏的视频文件。

（2）问题分析。

- 发现部分视频无法播放，可能是下载不完整或文件损坏。
- 标签中存在拼写错误或类别不一致。

（3）改进方案。

- 重新下载：对损坏的视频重新下载或从备份中恢复。
- 标签校正：统一标签命名，修正错误。
- 格式转换：将视频转换为统一的格式和参数。

（4）报告撰写。

- 概述：简述数据集的基本情况和检查目的。
- 发现：详细列出数据质量问题和对应的数据量。
- 建议：提出具体的改进措施和优先级。

（5）结果。

- 提高数据质量：经过改进，数据集更加统一和完整，为模型训练提供了可靠的数据基础。
- 经验总结：数据质量评估是数据使用的重要环节，应在项目初期进行，避免后续问题。

电子工业出版社有限公司 版权所有

3.4 开源数据在数据产品生产中的应用

3.4.1 利用开源数据扩充数据集

通过将开源数据与自有数据进行整合可以扩充数据集，数据融合时需注意数据分布不匹配、类别定义差异和数据格式不一致等问题，避免影响模型性能。

1．数据融合技术

（1）目的。

将开源数据与自有数据有效结合，扩大数据规模，提高模型的泛化能力。

（2）方法。

- 数据合并：直接将开源数据与自有数据合并，形成更大的数据集。
- 数据映射：对不同数据集的标签和格式进行映射，确保一致性。
- 特征融合：在特征层面对数据进行融合，提取共同的特征空间。

（3）注意事项。

- 数据兼容性：确保数据格式、分布、标注的一致性。
- 数据质量控制：对开源数据进行筛选和清洗，保持整体数据质量。

2．案例分析：通过数据扩充提高模型的泛化能力

某医疗影像分析项目，自有的数据集规模有限，难以训练出高泛化能力的模型。解决方案是引入开源数据，即从公开的医疗影像数据集（如 NIH Chest X-ray Dataset）中获取更多的影像数据。

（1）数据融合。

- 格式统一：将所有影像数据转换为相同的分辨率和文件格式。
- 标签映射：统一疾病标签，处理不同数据集中标签的不一致问题。

（2）模型训练。

使用融合后的数据集重新训练模型。

（3）结果。

- 模型性能提升：在独立测试集上，模型的准确率和召回率都有明显提高。
- 经验总结：数据扩充能够有效缓解数据稀缺的问题，提高模型的鲁棒性。

3.4.2　处理开源数据的挑战

开源数据通常存在数据噪声、缺失等问题，可使用半监督学习或主动学习方法处理开源数据中的噪声标注或缺失标注等问题，提升数据质量。

1．数据偏差与公平性问题

开源数据集可能存在种族、性别、年龄等方面的偏差，导致模型在特定群体上表现不佳；模型可能出现歧视性决策，影响公平性和可靠性。

2．案例分析：开源数据集中的偏差如何影响模型决策

在面部识别系统中，使用了某开源人脸数据集进行训练，但模型在不同种族人群上的识别准确率存在显著差异。原因是数据集中某些种族的人脸样本明显不足，导致模型对这些群体的识别能力不足，从而导致模型在实际应用中对少数族裔人群的识别错误率较高，可能引发歧视和法律问题。

（1）解决方案。

- 数据平衡：收集更多来自少数族裔的样本，平衡数据集。
- 公平性评估：引入公平性指标，评估模型在不同群体上的表现。
- 模型改进：采用公平学习算法，减轻偏差的影响。

（2）结果。

- 模型性能改善：在各群体上的识别准确率趋于一致，减少了偏差。
- 社会影响：提高了模型的公平性，避免了潜在的伦理和法律风险。

开源数据为人工智能的发展提供了宝贵的资源，但在使用过程中需要注意数据质量、许可协议和潜在的偏差问题。通过合理地获取、预处理和融合开源数据，能够有效地提升数据产品的生产规模、质量和进度。在实际应用中，应结合项目需求，谨慎地选择和处理开源数据，为数据产品的成功奠定坚实的基础。

电子工业出版社有限公司 版权所有

3.5 习题

1．操作题

选择一个开源数据集（如 IMDb），使用 Python 完成数据的预处理（文本分词、去停用词），展示处理前后的数据示例。

2．概念题

简述开源数据的三大优势和三大局限性，说明其对中小企业开发 AI 大模型的意义。

3．分析题

比较 CC BY 和 CC BY-SA 许可协议的异同，分析违反 CC BY-SA 协议的潜在法律后果。

4．实践题

使用 Kaggle API 下载“zillow/zecon”数据集，编写 Python 脚本统计“RegionName”唯一值数量，并使用 SHA-256 验证数据完整性。

5．设计题

为一个医疗影像分析项目设计开源数据融合方案，结合 NIH Chest X-ray Dataset 和自有数据，明确格式统一和标签映射的步骤。

6．案例分析

针对第 3 章中 ImageNet 推动深度学习的案例，分析其标准化基准如何促进模型创新，提出两种改进 ImageNet 偏差的方法。

电子工业出版社有限公司
版权所有