

面向大模型训练的数据集质量评估模型与方法研究

随着人工智能大模型技术的发展，制约大模型能力的因素越来越集中地体现在训练数据的质量上，这已经在行业内形成了广泛共识。如何评估用于大模型训练的数据集质量，成为业界的一个难题。

文 | 李哲洙 李洪金 砺英数智（北京）数据技术有限公司

一、概述

高质量数据对大模型训练的重要性已成为业界共识。高质量的数据不但能够确保模型学习到正确的特征和模式，避免模型出现偏差或过度拟合的情况，而且能够提高模型的准确性和泛化能力，使模型在面对未知数据时表现得更加稳定。因此，研究如何评估面向大模型训练的数据质量，对于确保大模型训练的有效性和成功至关重要。

二、大模型训练数据生态系统

评估数据质量的前提是要先明确评

估“什么数据”的质量，这是质量评估的基础。

数据作为生产资料，按照数据处理过程，我们将数据划分为原始数据、归一化数据、预训练数据、微调数据和价值观对齐数据，这些数据共同构成了大模型训练和优化的数据生态系统。将数据划分为不同类型是非常有意义的。因为不同的数据在大模型训练的不同阶段扮演着不同的角色，不同数据的具体质量评估指标因此也是有所不同的。

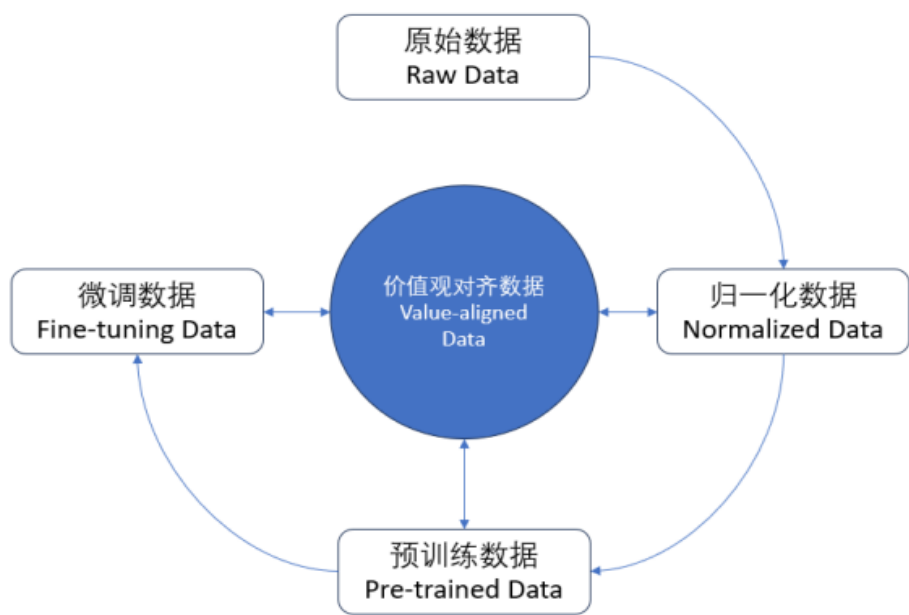
数据生态系统各个数据集的内涵与外延如下：



赛迪网官方微信



数字经济官方微信



来源：砺英数智（北京）数据技术有限公司

图 1 大模型训练数据生态系统

原始数据（Raw Data）：是指从现实世界收集来的、未经处理的数据。这些数据可能包含噪声、错误或不相关的信息，可能是结构化的（如数据库中的数据），也可能是非结构化的（如文本、图像、音频等）。原始数据包含所有可用的信息，是后续所有处理步骤的基础。

归一化数据（Normalized Data）：是指从原始数据中经过清洗、格式化、归一化等处理后得到的数据。这个过程去除了原始数据中的噪声和异常值，将数据的特征值调整到同一数量级或范围，消除不同特征之间的尺度差异。

预训练数据（Pre-trained Data）：是指用于模型预训练的数据。他包含模型所需学习的通用知识和结构。预训练的的目的是让模型学习数据的通用表示或

特征，这些通用表示可以捕捉数据中的统计结构和语义信息，使得模型能够更好地理解和处理输入数据。

微调数据（Fine-tuning Data）：是指在预训练模型基础上，针对特定任务进行调整和优化所使用的数据。这些数据通常包含了与目标任务更相关的信息，帮助大模型更好地适应特定任务或领域。

价值观对齐数据（Value-aligned Data）：这是一种特殊类型的数据，是指用于确保模型在训练和推理过程中遵循特定的价值观或道德准则的数据。这种数据包含有关偏见、公平性和道德标准的信息，用于在模型训练过程中进行约束或优化。

这些数据之间的关系是层层递进和

相互补充的。原始数据经过处理和转换后形成归一化数据，归一化数据用于模型预训练得到预训练数据。然后，使用微调数据对预训练后的模型进行微调以适应特定任务，在整个过程中，价值观对齐数据可以作为约束条件或优化目标来确保模型的道德和伦理表现。

三、数据质量评估框架

为了确保数据生态系统中，不同数据类型能在模型训练的不同阶段发挥出应有作用，确保模型不仅在技术上表现出色，而且在价值导向上也符合预期，需要对数据生态系统中的各个数据类型进行全面的质量评估。同时，在进行质量评估时，也要充分考虑到不同业务领域

的特点和要求。

因此，为了系统地评估数据在不同业务领域中，针对不同数据类型，采用的质量评估方法的不同，设计了一个多维的数据质量评估框架，它涵盖数据类型、质量评估层次和业务领域三个维度。数据质量评估框架如图 2。

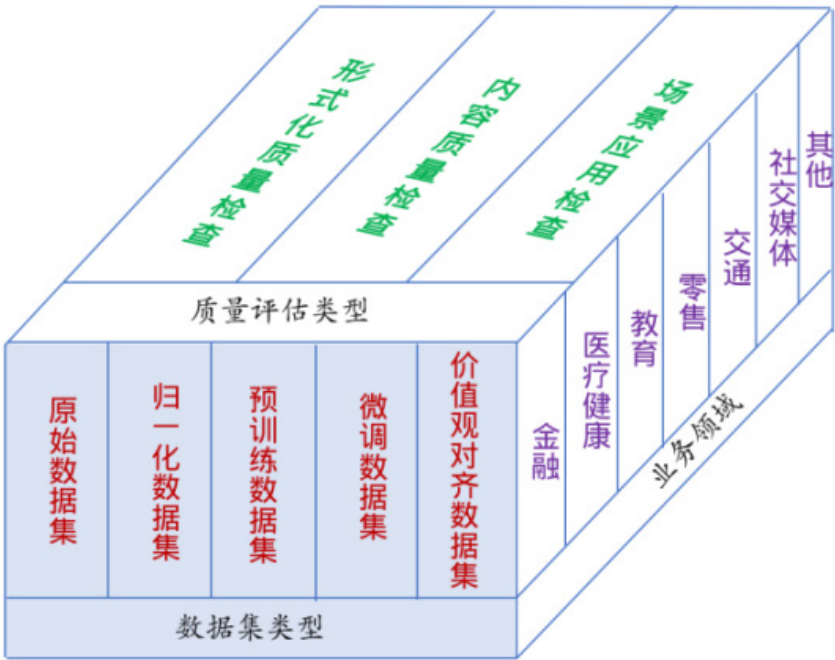
数据质量评估框架模型中：

数据类型作为模型的行，每个数据类型都需要全面的质量评估。

数据类型有原始数据集、归一化数据集、预训练数据集、微调数据集和价值对齐数据集等。

质量评估作为模型的列，每个质量评估类型都包含一系列具体的评估指标。

质量评估分为形式化质量检查、内容质量检查和场景应用检查



内容质量检查和场景应用检查三种类型：形式化检查主要是检查数据的结构、格式和完整性等，确保数据在物理层面没有问题；内容质量检查是评估数据中数据的准确性、一致性和可靠性等，确保数据在业务层面是可衡量的；场景应用检查是评估

来源：砺英数智（北京）数据技术有限公司

图 2 数据质量评估框架

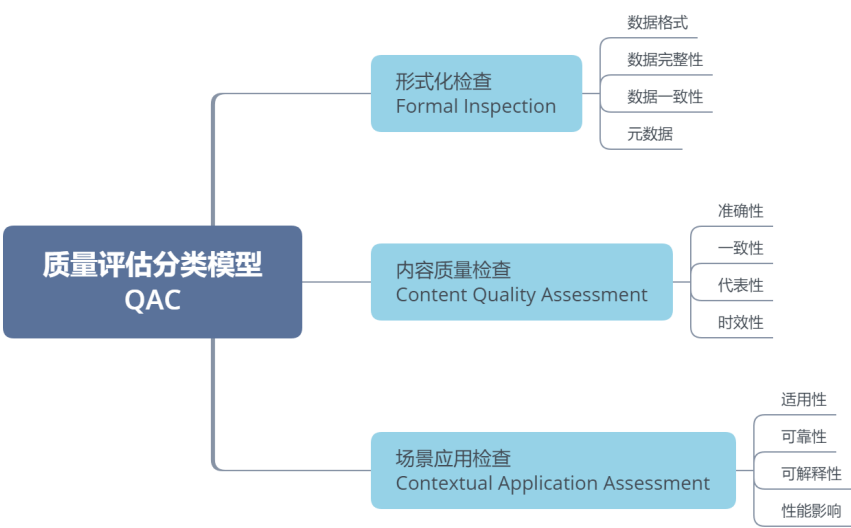
数据集在特定场景下的适用性，确保数据在应用场景下的可视化和客观评判。

业务领域作为模型的深度，每个业务领域都有其特定的需求和考虑因素。

不同垂直业务领域如金融、医疗、教

育等，在设计具体评估指标时需要进一步细化和调整。例如，在金融领域，可能需要特别关注数据的准确性和可靠性；在医疗领域，可能需要更加严格地遵守伦理规范和法律法规要求。

数据生态系统质量评估框架是一个多层次、多维的质量评估框架，包括数据类型维、质量评估维和业务领域维。针对每个“小立方”，可以定义明确具体的评估指标、评估方法或标准，并根据需要进行定制和调整。最终，通过将不同层次的评估结果进行综合分析和评估，得出整体的数据质量评估结果。因此，数据质量评估框架可以帮助我们全面、系统、定量化评估不同类型数据在不同业务应用场景下的质量表现。从而，为如何评估数据的高质量提供了总体框架和路径。



来源：砺英数智（北京）数据技术有限公司

图 3 质量评估分类模型

四、质量评估层次模型

为了对数据生态系统进行更加全面、系统和定量化的质量评估，在数据质量评估框架中，针对质量评估维，划分为形式化质量检查、内容质量检查和场景应用检查三类，形成质量评估分类模型（Quality Assessment Classification Model，简称 QAC）。分别从物理、业务和场景等不同层面对数据质量进行精准评估，为数据质量定量化评判提供指导和方向。

质量评估分类模型（QAC 模型），每个质量评估类型的目的如下：

形式化检查（Formal Inspection）：形式化检查主要关注数据的格式、结构和组织等物理特性。目的是确保数据在技术上是无误的，并且可以被机器学习算法或模型轻松地解析和使用。具体评估指标可以包括：

数据格式：检查数据是否遵循预定义的格式或标准，如 CSV、JSON、XML 等。

数据完整性：确认数据中没有缺失的值或损坏的记录。

数据一致性：确保数据中字段和类型是一致的，例如日期格式、数值范围等。

元数据：检查元数据是否完整，包括数据来源、收集日期、数据版本等。

内容质量检查（Content Quality Assessment）：内容质量检查关注数据的内容是否准确、相关和可靠等业务特性。目的是确保数据在内容上与业务高度契合，在业务覆盖度和深度等方面，能为模型训练提供高价值的信息。具体评估指标可以包括：

准确性：验证数据的信息在业务层面是否准确无误，特别是对于标注数据，标注的准确性至关重要。

一致性：检查数据中是否存在矛盾或重复的信息。

代表性：评估数据是否覆盖了所有相关业务的类别和情况，避免偏差。

时效性：对于需要反映最新情况的数据，检查数据的时效性和更新频率。

场景应用检查（Contextual Application Assessment）：场景应用检查关注数据是否适合特定的应用场景和业务需求。目的是确保数据在特定的应用场景中能够发挥预期的效果。具体评估指标可以包括：

适用性：评估数据是否包含了适用于特定任务的特征和属性。

可靠性：检查数据在特定场景下的稳定性和重复性。

可解释性：确保数据中的特征和标注可以被人理解与解释。

性能影响：分析数据对模型性能的潜在影响，包括训练效率、泛化能力等。

通过数据质量分类评估，可以更全面、更精准地评判数据的质量，确保他们能够满足模型训练和实际应用的需求。这种分类评估方法有助于识别数据的问题和潜在的风险，从而提高数据的可靠性和模型性能。因此，数据质量评估分类模型为数据质量评估提供了全面、系统和定量化的方法论，为数据分析和质量分析提供了有力支撑和保障。

五、总结

本文对大模型训练和应用所涉及的数据质量评估进行了研究。首先，提出了数据生态系统的概念，解决了质量评估对象不具体、不清晰的问题；其次，融入业务领域垂直大模型的典型特点，研究设计了质量评估框架，通过业务领域维、数据类型维和质量评估维的构建，为如何评估数据质量提供了总体框架；最后，针对质量评估维，研究设计了质量评估分类模型（QAC 模型），为数据质量全面、系统和定量化的评判提供了方法论。

责任编辑：杜玢翰 dubh@staff.ccidnet.com