从大模型与数据的关系看合成数据的价值

大模型技术在数据积累、算力支撑、模型精进及应用拓展四大维度上，均实现了显著突破。其中，合成数据的应用有效克服了现实世界数据在获取难度、规模限制及多样性不足等方面的挑战。本文将深入剖析大模型与数据之间的关系，着重探讨合成数据的独特价值，并进一步揭示大模型与合成数据在推动数智融合过程中所扮演的基石角色。

文丨**武朝尉** 北京市经济信息中心正高级工程师
徐玮彤 天云融创数据科技（北京）有限公司市场总监

一、合成数据，机器学习的未来

OpenAI 在其官方邮件中明确表示，自 2024 年 7 月 9 日起，将开始阻止来自非支持国家和地区的 API（应用程序接口）流量。为了维护服务质量和安全性，OpenAI 将采取额外措施来限制来自当前不支持的国家和地区的 API 流量。

这一举措对国内基于 OpenAI 大模型构建的应用生态无疑是一记重击。两个世界的技术正在逐步脱钩，越来越多的商业链条将被切断。这不仅影响物理世界，也波及数字世界。与此同时，市场上还存在另一种声音，认为这是一个巨大的机遇，可以借此机会加速国内搜索引擎等技术的发展。

从“深蓝”到 AlphaGo，再到今天的 ChatGPT，人工智能经历了从符号主义的知识封装，到连接主义的知识学习，再到当今生成式泛化表达能力的发展历程，并逐渐开始广泛应用于生产实践中。OpenAI 的 ChatGPT 作为人工智能的明星项目，市场上有太多追捧声音，过度神秘化 ChatGPT 并自我矮化，加上 OpenAI 宣布终止对中国提供 API 服务，国内技术的失败感扑面而来。

其实在北美市场，基础大模型的发展已经从单一模型训练 Trainging 发展为 Serving 多元化工业化基础设施化。

中国人工智能的发展如何摆脱依赖、实现优势破局和定义自己的生态？这需要找到我们技术的独特演进路径。与其将人工智能押宝在大模型上，不如重点看合成数据价值。

合成数据（Synthetic Data）是指通过计算机算法生成的模拟数据，他模拟真实世界的数据分布和特征，通过数

学模型和生成技术，来构建新的数据集，而不是直接来自现实世界的观测或记录。早在1993年，著名统计学家Donald Rubin就在论文中提出了合成数据的概念。近年来，随着ChatGPT的火爆和生成式人工智能技术的发展，合成数据概念受到越来越多的关注。

大模型训练数据通常来自网络获取数据、外部付费/开源数据集、企业自有数据以及AI合成数据。大模型训练和微调所需数据量快速增长，真实世界数据将在数年内被用尽。研究机构Epoch估计，机器学习可能会在2026年前耗尽所有“高质量语言数据”。据Gartner预测，2024年用于训练AI的数据中有60%将是合成数据。以Meta2024年7月发布的LLaMA3.1模型为例，监督微调环节的数据里有相当比例是合成数据，使用合成数据确实带来了模型效果的提升。

合成数据具有以下几个主要特点：

模拟真实数据分布：合成数据通过算法和模型模拟真实世界的数据分布与特征，使得生成的数据在统计特性上与真实数据保持一致。

保护隐私：由于合成数据并非直接来源于真实世界的观测或记录，因此可以在不暴露个人隐私的前提下，提供与真实数据相似的数据资源。

增强数据多样性：合成数据可以模拟和生成现实世界中难以采集到的边缘场景，增加数据的多样性，提升模型的泛化能力。

经济高效：通过算法生成合成数据，可以节省大量时间和成本，提高数据获取的效率。

合成数据提供了一种更快捷、更有效的方式来获取我们需要的数据，成本比从现实世界获取数据的成本更低，同时减少了烦人的数据隐私问题。高质量的真实数据已逐渐无法满足大模型训练与精细微调的需要，这就促使合成数据作为真实数据的重要补充，在人工智能领域扮演着日益关键的角色。合成数据作为算法、生成模型及模拟技术的产物，能够模仿现实世界数据的特征与模式，为大模型的训练与优化提供丰富的数据资源，正在成为机器学习的未来。

二、大模型使用数据的方法

知己知彼，百战不殆。目前人工智能市场有两条核心路径来使用数据，其使用方法分别是：

第一条路径是无条件地依赖规模法则（scaling law）。相信只要把数据喂进去就会涌现机制，用算力和数据堆积给已有的算法实践Transformer，这种路径就是在已知的知识结构里去寻找未知拼接的方法，能力是有限的，就像在陆地上看着教科书学游泳一样，一定会遇到模型基础理论的天花板。在已知中组合筛选（条件概率）获得的知识，只能是补齐现有的知识拼图，例如通过大量的实验发现新的元素，可以补全门捷列夫发现的元素周期表，但是不能诞生量子力学对基本粒子的理论和元素生成公式，更不可能产生牛顿的“加速度”、阿拉伯数字“零和无穷大”这样的革命性知识。从学外语到学母语，从建立认

知再到推理和逻辑，其路径完全不同。

第二条路径不再依赖单一大模型而是和更多的系统架构配合来组成一个务实的AGI工程架构。谷歌、微软、亚马逊、HuggingFace都选择了这条路径。谷歌发布的大模型成熟度参考架构，定义了从L0到L6的分级，调用GPT直接使用单一大模型的能力仅仅是L0水平。加入提示词工程、精调模型Lora的意图理解、向量数据库寻回私域数据的记忆、Agent规划拆解、plugin执行和反思等，可逐步完善大模型成熟度到更高等级。

国内能同时提供数据供给侧的混合负载数据库和数据消费侧的机器学习平台的天云数据公司CEO雷涛表示："关于大模型与数据之间的关系，核心是存量数据和增量数据。关于存量数据，核心需要关注的技术是大模型to DB，去解决如何跟上万张表且高价值密度的企业数据库的数据发生关系；关于增量数据，去解决如何持续地供给大模型以及大模型真正的算力出口在哪里，是提供服务还是提供新兴的生产资料。供给更多的数据资源，也就是合成数据的概念。"

存量数据是这两年的主流，就是把已有的知识做知识封装和知识移动，是一种端到端的训练方法；增量是用RAG、向量数据库外挂在模型之上，将增量的信息全部训练进去。

增量的一个核心关键词是合成数据，供给大模型的数据资源从哪里来？这里面涉及非常多的场景，最早使用合成数据更多是面向专业领域的大模型微调，需要有非常精准的且合适的数据才能提供准确的大模型服务。Lora是一种常见的微调方法，但他对输入给模型的数据的要求也非常高。市场如何获取这种数据？比如做一个客服系统，每家企业都有各自的产品手册、规章制度，但是客户会提什么样的问题呢？传统的方法是用人工标注采集的方式去获取这些Q&A，现在市场完全可以针对产品手册的大模型来生成Q&A，这就是典型业务场景的合成数据。

合成数据已经开始从模型训练数据的生成到直接场景数据生成，大模型进入到了数据飞轮效应，就像Robot让我们看到的里程碑式技术是机器在供给自己。可以简单类比理解一下，就是供给给机器训练所需要的数据就像汽车要加的油一样，开始变成是自己生产出来的。

客观来说，针对驾驶而言，一些极端灾害、路况交通事故是不可能通过大规模的路面采集获取的。这种数据集叫CoreData，是可以通过模型来生产的，在项目上需要交付的就是合成数据集。

合成数据的生成方法多种多样，常用的有以下几种：

1. 基于LLMs生成的合成数据

LLMs（Large Language Models）拥有卓越的语言理解和表达能力，以及强大的指令遵循能力，能够为特定场景和领域创建定制的数据集。使用LLMs生成合成数据的常见做法，可分为提示工程和多步骤生成。

提示工程：基于高性能模型的提示工程生成合成数据，用于补充特定领域的数据，帮助轻量级或下一代模型进行

监督学习。例如，Meta 的 LLaMA3 模型在后训练阶段完全依赖于从 LLaMA2 获得的合成数据；OpenAI 计划利用其 o1 模型生成合成数据，以训练即将推出的 Orion 模型。

多步骤生成：利用模型生成的多步骤合成数据可以补充思维链（Chain of Thought，CoT）的中间推理步骤，有助于模型的对齐和进化。例如，浙江大学和中国科学院等研究机构使用 GPT-4-Turbo 模型生成代码和图像，并逐步引导模型生成解释答案的原理，构建起多模态合成数据集。利用这些数据集对 Vanilla Llava-1.5-7B 模型进行微调，可以显著提升其视觉推理能力。

2. 基于生成对抗网络（GANs）或扩散模型（Diffusion Models）生成的合成数据

生成对抗网络（GANs）和扩散模型（Diffusion Models）通过对抗训练和逐步去噪的方法，能够产生与真实数据高度相似的合成图像样本，广泛应用于数据增强、医疗隐私等领域。

3. 基于统计和模拟生成的合成数据

通过观察真实的统计分布，利用算法生成符合特定统计分布的数据；或者通过模拟器等方法创建数据，如 Sora 文生视频模型用到 Unity、UnrealEngine 等游戏引擎合成的视频数据作为训练集。在实际应用中，多种方式往往相互协同和补充，以提升数据合成质量。

大模型训练通常需要大量的数据。这些数据往往存储在各种数据库中。数据库提供了结构化和非结构化的数据源，供大模型在训练过程中使用。数据库系统能够高效地存储、检索和管理大量数据，使得大模型能够从中获取所需的信息。例如，训练语言模型时，数据库可以存储大规模的文本数据，方便模型进行访问和处理。

大模型需要连接价值密度高、逻辑性强、动态且鲜活的数据，这些数据都跟生产经营的交易相关，比如股票信息、金融账户、医院挂号信息。我们知道这些数据都不在静态的文档、文献或报告里，而是在数据库里。但是目前大模型所依赖的数据资源局限于静态文献中的知识，这在一定程度上限制了其对于高价值数据的全面获取，尤其是那些存储在客户私域中的宝贵数据。目前普遍采用的 RAG 技术将信息检索和生成两个阶段结合起来，通过检索数据库中的相关信息来辅助生成过程，解决大模型数据滞后带来的幻觉问题，提高生成内容的质量。

供给大模型的数据资源从哪里来？这里面涉及非常多的场景，最早使用合成数据更多的是面向专业领域的大模型微调，需要有非常精准的且合适的数据才能提供准确的大模型服务。

当 LLM 面临无法获得大规模、多样化标注数据集时，应该如何破解？英伟达 Nemotron-4 340B 构建了一个高质量合成数据生成的完整流程。值得一提的是，指令模型的训练是在 98% 的合成数据上完成的。Anthropic 使用一些合成数据来训练其旗舰模型之一——Claude 3.5 Sonnet，Meta 使用人工智能生成的

数据微调了 Llama 3.1 模型，OpenAI 正在从 o1 模型中获取合成训练数据用于即将推出的 Orion 模型。

这有可能彻底改变训练 LLM 的方式。选择合适的方法和技术，可以显著提高模型的性能和泛化能力。未来，各行各业都无须依赖大量昂贵的真实世界数据集，用合成数据，就可以创建性能强大的特定领域大语言模型。

三、合成数据带来数据飞轮效应

合成数据的应用，不仅有效克服了现实世界数据在获取难度、规模限制及多样性不足等方面的挑战，更为开发出更加健壮、可靠且公平的大模型奠定了坚实基础。具体而言，合成数据尤其适用于那些数据稀缺或难以直接获取的特定领域。此外，合成数据还能根据具体需求进行定制化设计，如确保不同类别数据的平衡表示，进一步提升模型的泛化能力。同时，合成数据还有助于缓解数据隐私保护的压力，通过创建匿名化或去标识化的数据集，为数据的安全共享与高效利用提供了保障。

比如在自动驾驶领域，NVIDIA 利用合成数据来增强自动驾驶汽车的摄像感知系统。他们通过 NVIDIA DRIVE Sim 仿真平台生成远场物体的合成真值数据，并将这些数据添加到现有的真实数据集中，以训练可探测远距离汽车的网络。这种方法显著提高了自动驾驶汽车对远场物体的感知能力。

比如在金融领域，天云数据服务于券商的数字人，播报的内容是来自于实时交易系统的数据和研报文本内容的结合。针对这样的场景，需要把大模型的模糊意图匹配和精确的 SQL 操作形成连接。这种连接不是一对一的，涉及非常复杂的工程技术。如何保证像 ASR 语音识别这些机器学习模型回答一个准确的答案？比如现在的销量是多少，是产品的销量还是区域的销量。像这样模糊的意图匹配，怎么和数据库里精确的字段完成匹配？在后台，需要准备大量的密集计算操作。传统的 MPP 数据库是没有并发能力的，可能只能支撑一句话十几个 token 的内容。但高并发任务，成百上千个宽表的 OLAP 执行对数据基础设施的要求非常高，只有 HTAP 数据库能胜任这种大模型的高并发 AP 类业务的底座。

在大数据和 AI 时代，数据不再只是“支持性资源”，而成了许多企业的核心推动力。谁能将数据用到极致，谁就能在市场中夺得先机。合成数据带来“数据飞轮”效应，将海量数据转化成实实在在的业务增长引擎，让企业的数据从“成本中心”变成了“利润中心”，在积累数据的过程中逐步提高运营效率、优化用户体验，形成了一种不可替代的增长模式。

责任编辑：杜玢翰　投稿邮箱 zhouhl@staff.ccidnet.com