大模型训练数据白皮书

LARGE LANGUAGE MODEL'S TRAINING DATA WHITE PAPER

Alibaba Group 阿里巴巴集团 ADC 数字中国研究院 阿里云

大模型训练数据白皮书

专家委员会

安筱鹏 阿里云智能副总裁

袁　媛 阿里研究院执行副院长

宋志刚 数字中国研究院（福建）院长

编写组主要成员

傅宏宇　王　峥　赵志远　郑达真

张　荣　陈岳峰　李天宇　徐　强

编写单位

阿里巴巴集团

数字中国研究院（福建）

阿里云智能集团

欢迎关注“阿里研究院”公众号

AliResearch
阿里研究院

关于我们

阿里研究院是阿里巴巴集团的内设智库机构，多年来扎根于阿里巴巴丰富的数字科技商业生态，依托前沿的产业实践和大量的创新案例，围绕集团“用户为先，AI 驱动”的战略重心，聚焦于科技创新、数据和算法治理等领域的研究。

联系我们

aliresearch2023@service.alibaba.com

目录

CONTENTS

自《中共中央国务院关于构建数据基础制度更好发挥数据要素作用的意见》发布以来，我国数据要素建设不断深入，在国家数据局等 17 部门联合印发的《“数据要素 ×” 三年行动计划（2024 - 2026 年）》进一步明确“建设高质量语料库和基础科学数据集，支持开展人工智能大模型开发和训练”。通过数据要素建设推动人工智能大模型发展，可以有效解决我国人工智能，特别是大模型研发所面临的数据瓶颈，进一步发挥大模型对于世界知识数据的汇集和处理能力，创造更大的生产力，助力我国从数据经济走向智能经济新发展模式。

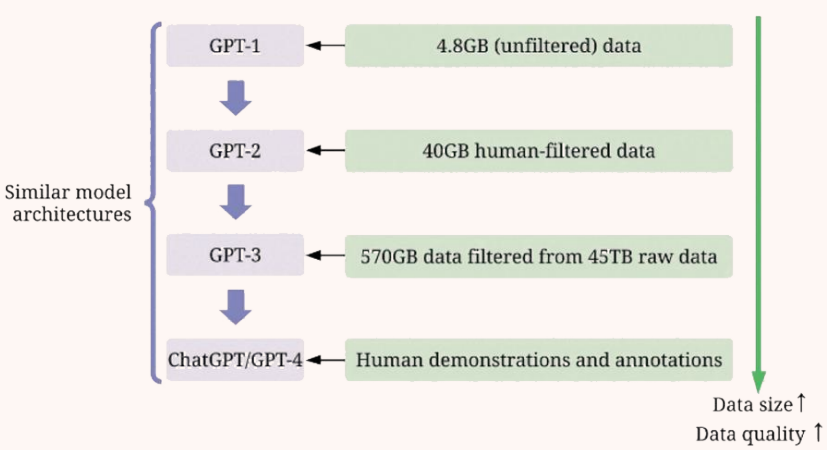
大模型是数据要素价值释放的最短路径，通过理解其训练所使用的数据类型，可以更好理解大模型发挥价值的内在机制，破解对训练数据常见的迷思和误解。而促进高质量训练数据的建设，需要理解人工智能对数据的实际需求，科学评价数据的规模和质量；需要综合利用政府、企业、社会等各方资源，构建共享、共创、共赢的合作生态，以更务实、多元、开放的方式解决供给不足的问题；还需要为技术发展预留空间，构建更顺应模型发展的数据治理体系，相信随着技术的日益成熟，相应的商业模式和制度设计也都会逐步完善。

01

训练数据对大模型发展的重要性

业界认为，算法、算力与数据，是支撑大模型发展的三大基石。更高质量、更丰富的数据是以 GPT 为例的生成式人工智能大模型成功的驱动力。GPT 模型架构从第 1 代到第 4 代均较为相似，而用来训练数据的数据规模和质量却有很大的不同。GPT-1 是由 4.8G 未过滤原始数据训练，GPT-2 是由经人类过滤后的 40G 数据训练，GPT-3 是由从 45T 原始数据中过滤的 570G 数据训练，而 chatGPT/GPT-4 则是在该基础上又加入了高质量人类标注。以吴恩达（Andrew Ng）为代表的学者观点认为，人工智能是以数据为中心的，而不是以模型为中心。“有标注的高质量数据才能释放人工智能的价值，如果业界将更多精力放在数据质量上，人工智能的发展会更快”。

数据在AI模型中发挥的重要作用-以*ChatGPT*为例



更高质量、更丰富的训练数据是GPT模型成功的驱动力

资料来源：Data-centric Artificial Intelligence: A Survey

02

模型训练所需的数据类型

数据作为大模型训练的基础，它提供了大模型所必需的知识和信息。区别于以往搜索系统、个性化推荐等所需的大量用户行为和偏好数据，随着技术的演进，大模型所需的数据是对知识性内容有强需求，是一种新的类型。

2.1 训练大语言模型的数据

大模型所需要的数据根据训练的阶段有所不同。以 ChatGPT 为代表的大语言模型（LLM）为例，其训练过程分为预训练（Pre-training）、监督微调（SFT）、基于人类反馈的强化学习（RLHF）三个阶段，后两部分又统称为

“对齐”（Alignment）阶段。

第一阶段预训练所需的语料是各种类型的世界知识，包括网页、书籍、新闻、论文期刊、对话文本、代码等形式，通过大量学习世界知识，构建模型的基础能力，理解客观世界的规律，该阶段的语料特征可以概括为“广”。

第二阶段 SFT，通过标注人员设计问答，编写正确答案，将例题投喂给模型，并希望模型在没有见过的任务中“举一反三”，提升泛化能力。第三阶段 RLHF，训练目标是让模型的价值观与人类对齐，需要人类对模型的回答进行打分、排序，让模型知道 " 怎么说更好 "。第二和第三阶段的数据质量要求较高，需要来自人类的高质量反馈，语料特征可以概括为“齐”。

如果将模型微调后部署应用于特定的场景形成行业大模型（如工业、金融、医疗等），则需要满足该场景专业需求的特定领域知识做预训练和对齐，需要具备一定专业深度，如行业数据库、专业文档、专业网站等，这部分的语料特征是“专”。

2.2 训练多模态模型的数据

大语言模型迅速发展的同时，Transformer 开始迁移到图像、视频和语音等其他模态数据领域，并与大语言模型融合，形成多模态大模型。多模态模型模拟人类大脑处理信息的方式，把各种感知模态结合起来，以更全面、综合的方式理解和生成信息，最终实现更丰富的任务和应用。从以 Mid-journey 和 Sora 为例的多模态大模型看，在训练阶段需要大量图像 - 文本对、视频 - 文本对等有标注数据集进行训练。图像 - 文本对是包含一张图像和一段描述该图像内容的文本的数据，让模型学习组成图像的像素之间、文字与图像的关联。视频 - 文本对包括一个短视频和一段描述视频中发生事件的文本，让模型不仅学习单个画面，还需要理解视频中的时间序列和动态变化。

2.3 训练数据的常见疑问和误解

2.3.1 大模型训练并不依赖用户个人信息

人工智能经历了从有监督学习到无监督学习的发展阶段，神经网络等技术推动了数据驱动的应用模式。传统的决策类人工智能在需求侧通过学习和分析海量的用户行为数据，判断用户的偏好和需求。在供给侧通过学习内容的特征，借助推荐、排序等机制实现需求和内容的匹配，并根据用户的行为反馈进行优化，提高算法的准确性。以个性化搜索为例，以大量的用户使用记录、用户画像、内容画像等原始数据为基础，提炼出客群和内容标签等不同维

度的信息，进而抽象出特征向量，用向量的空间距离计算用户和内容的相似度，通过匹配与排名进行个性化的搜索结果召回。基于上述特点，此类决策式人工智能技术在需求侧需要更多用户数据，在供给侧依赖更为全面的内容特征。

与以前的决策类人工智能相比，以大模型为代表的生成式人工智能的技术特征有明显差异。大模型是模拟人类的思维活动方式生成人类可以理解和使用的内容，而训练数据也是基于世界知识，对语料库等知识性内容有强烈需求，因此大模型训练阶段不依赖个人信息等原始数据。此外，为保证生成内容与人类价值观对齐，业界往往利用强化学习，通过纳入人工标注等机制优化表达，使模型生成内容更接近于人类认知。因此大模型对于用户数据并不依赖，而对专业化、高质量语料的知识性内容依赖大。由此看出，随着技术的演进，对训练数据的需求类型也有所不同。

然而，有很多人对此仍存在误解。根据第三方专业机构测评显示，超过 60% 的受访者误选了“盗取、泄露个人隐私数据的安全风险”作为大模型的最主要风险点。与一般看法相反，过量的个人数据会负面影响大模型的能力，而过于个性化的应用也将增加大模型的运算负担。对此，OpenAI 负责人 Sam Altman 表示，ChatGPT 不需要用户的个人数据，用户可以选择删除其与 ChatGPT 的交互历史；类似的，我国目前主流大模型在提供用户隐私保护的基础上，并不过度收集和使用用户个人信息，并允许用户控制和删除其与大模型交互的对话和提供的内容。当然，在大模型的推理阶段，如果用户恶意诱导，尽管有相应的模型安全机制，仍不能完全避免个人信息泄露的问题。但可以明确的是，大模型在训练阶段并不依赖个人信息。

2.3.2 中文语料短缺不是制约我国大模型发展的重要因素

谈到中文大模型，一个普遍关注的问题是，中文语料和英文语料在互联网中的占比存在显著差异：在全球网站中，英文占 59.8%，而中文仅占 1.3%，那中文语料供给短缺是否是制约我国大模型发展的关键要素呢？在实践中发现，规模并不是决定性影响因素。一是世界知识的积累有的属于客观事实，用英文或中文表达，其原理是一致的。或者说，在机器翻译质量有保障的前提下，可以弥补这部分中文语料的缺少。二是在训练技术上引入新方法也可以弥补语料供给不足的问题。例如通过合理安排不同语言类型的训练顺序，也能让模型学习到供给相对较少语言的丰富特征。

然而有一种类型的中文语料是极为重要且存在短缺的 - 中式价值观类语料。因为模型为了更好地理解客观世界和掌握规律，需要学习大量来自知识和价值观层的数据，它们更多受到人类主观意志的影响。而大模型是概率分布模型，其使用的数据来源分布将使得模型具备与之相似的人类意志。所以，训练中加入更多代表中式价值观的语料，有助于大模型更好地理解和反映中文使用者的文化背景和价值取向，从而在全球化的背景下保持文化的多样性和独特性。而且此类语料短缺的问题也没有办法通过机器翻译弥补，因为即使翻译质量有保障，仍会引入源语言的偏见，体现的仍是源语言的价值观。总体来看，文言文、古汉语、电子书籍等反映优秀传统文化的内容，以及主流媒体发布的能反映本土价值观的内容，都可视为高质量具有中式价值观的语料。但目前看，与语料相关的各环节：

从积累机制、数字化（比如我国古籍数字化率不到 30%），到开放共享与开发利用，及训练过程中机器算法与编码系统的建设，都仍需大量持续投入精力。可见，中文语料“量”的短缺尚可有解决方案，但中式价值观类的语料短缺，则会成为制约我国大模型发展的短板。

03

科学理解高质量数据的含义与作用

在生成式人工智能时代，模型训练的成功与否与所依赖的数据质量息息相关。模型的能力很大程度上可以反映出其训练数据的质量，这也无疑凸显了高质量数据在大模型训练和应用中不可替代的重要性。

3.1 高质量数据的重要性

由于高质量数据可以更好地模拟客观世界，将其作为训练数据可以增强模型能力。从技术层面看，通常用损失函数来量化模型预测输出与实际目标之间的不匹配程度。能更好模拟客观世界的高质量数据，可以使模型预测的概率分布尽可能逼近实际数据的真实分布，通过优化算法调整模型参数，让模型在训练集上的损失函数最小。从模型能力表现看，一是高质量数据可以提升模型的准确性和稳定性。首先，这些数据通常包含更准确和丰富的信息，有助于模型更好地理解数据的内在结构，掌握世界规律，提升产出的精准性。其次，数据清洗是提高数据质量的重要环节，包括去重、删除个信隐私内容、纠正错误、填补缺失值等，经过清洗的数据可以提升训练阶段的稳定性。二是高质量数据具有多样性，可以降低模型对特定数据集的依赖，提升鲁棒性和泛化能力。一方面高质量数据通过对现有不同来源的数据加以混合，调试配比，提升模型执行下游任务的泛化能力。另一方面可以利用数据增强等手段有效提升多样性，即通过对现有数据进行变换或扩充，如旋转、缩放、亮度调整等，生成更多的训练样本，增加训练数据代表性和多样性。

然而，即使在训练各阶段中的语料都满足高质量，能做到“真实性”、“准确性”、“客观性”、“多样性”的要求，仍不能完全避免模型结果产生幻觉，即“一本正经胡说八道”。因为大模型本质是概率模型，是基于前文预测

下一个词出现的概率，“词语接龙”出现的下一个词并不是 100% 有确定性的。所以高质量的语料，可以大幅降低模型结果产生幻觉的概率，但并不能完全避免。

但如果在训练中使用了较多错误、有毒、重复的低质量数据，则会对模型能力产生破坏性影响，也就是人们常说的“Garbage In Garbage Out”，比如没有经过严格质量筛选的大量社交媒体对话、用户生成的内容等。在对模型能力的损害上，使用错误的数据进行训练，会导致模型记忆有偏差信息，发生事实性错误；使用有重复的语料，则可能会导致模型在训练过程中对特定类型的示例产生偏见，降低生成结果的多样性，造成模型能力的显著下降。

由于高质量数据如此重要，会引发对一系列问题的思考，比如我们能否前置制定统一的标准体系，把高质量训练数据先识别出来？数据质量与模型的能力有什么联系？

3.2 高质量数据的标准

3.2.1 高质量数据类型的三重不确定性

第一重不确定性来自于所需的语料种类，其类型是由人类对模型能力需求决定的，而能力需求又是根据需要模型所完成的任务而不断演变。回溯基础大模型的发展历程，在 2020 年左右，基于 Transformer 架构的 Google Meena，其目的是让模型具有生成连贯且有意义内容的对话能力，因此对话文本被视为最重要的高质量数据。而随着技术路线的演进，人们发现更通用的上下文理解是重点，因此书籍和科研论文等又被视为高质量数据。通过提升其在训练语料中的占比，可以增强模型从文本中捕捉长距离依赖的能力。随着人们对通用人工智能的向往，对提升通用性能的北极星指标 - 推理能力有帮助的语料，又更加被重视。一种是代码数据，因为里面涉及大量 If-Then-Else 等条件控制信息；另一种是教材，因为涉及了比较详细的数学推理过程，和逻辑链高度相关。如果再拓展到行业模型，根据对模型能力的不同需求，语料类型更难以一一列举。比如，经人类标注的，由视觉相似性图片构成的匹配对数据库，可以作为高质量数据用于大模型在广告领域的训练，通过更好预测用户需求实现对素材点击率的优化。而通过收集人类驾驶员对稀有事件（比如驾驶过程中遇到的复杂路况、极端天气、异常行为的人或车辆等场景）的应对数据，则可以更好训练完全自动驾驶（FSD）模型在不同场景中的处理能力。由此看出，由于生成式 AI 在技术演进和应用场景拓展中具有不确定性，模型对所需要语料类型也在发生变化，“高质量语料”的类型和范围也在不断拓展。

第二重不确定性来自于语料形态的演化，高质量数据的形态会不断增强，以强化该类型语料的能力。一方面随着合成数据和数据增强技术的提升，大模型正在不断拓展对数据利用的可能性。如领域知识生成，对于大模型难以直接使用的原始数据，通过加工、改造和泛化可以形成模型训练可用的知识类数据。另外，在自动驾驶等领域，通过仿真数据生成更多样化、不同视角的物理世界用于模型训练，可以提升针对特定场景的数据收集效率，弥补真实

世界中对稀有事件观测不足的问题。另一方面，随着模型长上下文建模能力的增强，对代码和教材的需求又有了质的变化。例如，训练用的代码数据从执行单一任务到仓库级，让模型推理能力从掌握单任务模块进化到学习整体架构；训练用的教材从中小学级别知识拓展到大学，进一步增强了复杂场景下的推理能力。

第三重不确定性来自于不同数据类型之间的有效搭配，数据调度对模型能力起到重要作用。该环节强调对不同来源的数据加以混合，以提升数据集的多样性。因为不同类型的数据对模型能力提升的侧重点不同，各个数据来源的配比不同，也会影响模型的泛化能力以及在下游任务的表现，其中包含两个重要环节：一是调整不同来源数据的配比（数据混合），二是不同来源数据用于训练的顺序（数据课程）。

数据混合环节可以在训练的不同阶段设定配比，在实践中不断尝试出最优的组合。例如在监督微调阶段，有研究者从 Stack Exchange、Reddit 等网站中精选高赞语料，配合手工整理的问答对，得到共计 1000 条高质量微调数据，以“少而精”的数据在模型对齐能力上取得了很好的效果。数据混合在实践中会采取不同策略，一是增加数据源的多样性，这对大模型在下游任务能力的提升十分关键；二是可以根据大模型执行的目标任务，选择对任务性能产生积极影响的数据。数据课程环节是为了让大模型更好地学习某项技能，对语料学习顺序进行探索。一般来说，按照技能集合的顺序组织预训练语料（从基础技能到目标技能），比直接从专注于目标技能的语料库中学习更为有效，如从通用或简单的例子开始，逐步引入更具专业化或复杂度的数据。

3.2.2 同类数据的评估标准并不完全一致

对同类语料的质量评估，往往从质量、规模、多样性三个维度出发。在质量上，被视为“高质量”通常是因为其信息已经通过了有用性或质量筛选，这些大多可以从来源中做判断。例如，在语言模型训练中，新闻、科研论文或开源代码项目中的内容会受到专业标准（如同行评审）的筛选；常识性内容中，维基百科则经受了一群专注编辑者的筛选；而经过筛选的对话内容则是基于用户的积极互动（如在 Reddit 上获得的点赞数量）；在多模态模型训练中，以视觉中国为例，其网站有经过专业设计师筛选的大量图片和视频素材，并有对图像的光照、构图、艺术性、美观性等专业性标注，形成了高质量的图像 / 视频 - 文本对。其次，对于无法从信息来源直接判断数据质量的语料，人们会尝试用评估模型进行打分。例如对大量公开的网页，通过先对少量样本人工评价得到可读性、帮助性、安全性等指标，通过这些具有代表性的样本训练评估模型，将人工定义的评价标准转化为机器可识别的特征和模式，在此基础上评价语料中所有网页信息的质量。然而，即使有了前两种方法，针对部分语料仍无法前置判断其质量。如用于领域模型训练的语料，涉及到不同行业的专业知识，缺少统一的判断标准，往往是在模型训练中不断检验其质量的高低。

从规模看，收集足够规模的高质量语料也非常重要。根据大模型“伸缩法则”，当模型的参数或计算量按比例扩大时，模型性能也与之成比例提升。而随着参数规模的增加，也需要更多数据来训练模型，即模型参数与训练语料之间也存在类似的比例关系。需要指出的是，并不是语料规模越大越好，而是高信息密度的语料规模越大越好：以 CC（Common Crawl）和 C4 数据集的对比为例，CC 是一个有 400TB 的公共网络抓取数据集，包含了互联

网上数十亿网页，内容非常广泛但未经清洗。而 C4 则是对 CC 进行了过滤噪声、重复内容等清洗后的 305GB 数据集。经评估发现基于 C4 训练的模型性能优于 CC，这既说明了数据清洗的重要性，也说明了语料规模不能一味追求大。

此外，同类型语料中的多样性也是值得关注的问题。首先，会涉及到数据集的公平性，从网络采集的信息存在对于弱势群体（如种族、性别、职业、年龄等）不平衡的问题，可能会加剧现有偏见或系统性不平等。在技术层面上，通过对训练数据集进行仔细地审查和筛选，确保其分布的广度和均衡性，可以缓解公平性问题。另外，同类语料的多样性也会影响模型能力，特别是在安全能力建设方面。真实世界中潜在隐患的出现往往是偶然事件，相较于对这些“不良信息”的一概删除，对这些样本采用打安全标签的方式，反而有助于提升模型对安全风险的识别，增强安全防护能力。

针对不同类型的高质量语料，意味着其在语料类型、语料形态以及语料搭配使用三个层面存在不确定性。而针对同类型的语料，又涉及到从质量、规模、多样性三方面的综合考量，对高质量并没有统一的评估标准。就像生成式人工智能技术的发展路径充满不确定性一样，对高质量数据的判断，也同样没有人拥有“上帝视角”，可以精准前置预知高质量的标准，来决定哪些是未来的高质量数据。

因此，在对高质量数据的理解上，应认识到对高质量并不适合被前置的客观标准定义。“高质量”更多是一种主观判断，它的标准取决于模型的应用目的，数据类型会根据模型的发展阶段“因时而动”、根据技术人员的理解判断“因人而异”、根据模型的训练效果“因效而定”。因此，所谓“高质量标准”的制定，至多也只是对同类型数据在质量维度评估提供一种参考，对模型训练的价值有限。

04

合成数据作为解决训练数据供给不足的新方案

4.1 训练数据供给不足带来的思考

在生成式人工智能技术不断发展的趋势下，训练数据来源是人们最关心的问题之一。上节以政府和社会力量的视角展开。本节以已经使用的数据源和正在探索的新数据源视角展开。在已经使用的训练语料中，有用于语言大模

型训练的文本数据，包括网页信息、书籍、科研论文、知识百科、专业问答、代码以及领域知识，也有用于多模态模型的图片、视频、音频等媒体数据。根据 Epoch AI 的估算，书籍、科研论文等高质量语言数据集可能会在 2024 年前耗尽。人们正在积极探索新数据源，以缓解训练语料可能面临不足的问题。一种思路是将未数字化的知识数字化，如在最新发布的 Claude 3 中，提到了将大量未数字化的书籍和资料做数字化处理，成为模型可读取的训练语料。还可利用机器感知数据，比如将无人车、无人机、其他智能硬件设备等生成的大量物理世界数据用于训练。另一种思路是利用模型或算法，批量生成新数据，比如合成数据，然后利用它们训练模型。

近期，合成数据在大模型训练和应用的话题引起了广泛关注。一方面，高质量的合成数据可以作为真实数据的补充和替代，模拟现实世界的复杂性和多样性，被视为扩展模型学习范围与能力的重要手段。另一方面，合成数据的生成过程可能存在偏差或噪声，导致其质量和真实性无法完全模拟客观世界。由此引出一系列值得深入讨论的问题：对于合成数据的价值，它能否拓展大模型能力的边界？又是否能替代真实数据，缓解优质数据供给不足的问题？此外，合成数据能否通过对现有数据的深加工，将之前不能被用于训练的数据转化为可用，提升模型对数据利用的可能性？而对于合成数据的风险，人们也会担忧是否会出现“大模型自己产生数据进行自我训练”的循环，导致初始偏差被不断放大，最终使模型失控？这种新数据源还会带来哪些新风险？

4.2 合成数据的定义

合成数据是通过算法和数学模型创建的。首先建模真实数据的分布，然后在该分布上进行采样，创建出新数据集，模拟真实数据中的统计模式和关系。合成数据类似于数据的“替身演员”，发挥补充或替代真实数据的作用。在机器学习和人工智能领域，合成数据可以为模型提供训练材料，帮助它们学习、理解和预测。需要注意的是，如果生成过程设计不当，合成数据也可能缺乏保真度，对客观世界的模拟出现偏差。

4.3 合成数据的必要性

什么情况下会用到合成数据？本质原因是真实世界中获取数据遇到困难。一是真实世界中难以观测，如罕见病或极端天气等。利用合成数据可以设计比真实数据集更广泛的情况，对 Corner Case 进行模拟，提升训练数据集的全面性和多样性，确保在处理边缘案例时也有良好性能，提升模型泛化能力。二是真实世界中数据获取的成本高，如大模型对齐训练中需要人类大量的高质量反馈。利用合成数据可以实现对齐流程自动化，几乎不需人类标注，大幅节省成本，提高获取效率。三是数据获取和处理涉及到真实世界中的个信甚至敏感信息，特别是医疗健康

和金融领域。合成数据可以利用差分隐私对个体信息“加噪声”等方法，模拟真实数据集的分布，而不模拟其中的真实个人信息，实现对个信去标识化。由此归纳出，合成数据具有全面性和多样性、经济高效、有利于隐私保护等优点。

4.4 合成数据的生成方法及分类

根据是否基于实际数据集生成，合成数据生成方法主要分为两大类。第一种是基于真实数据集构建的：人们会建立模型以捕获真实数据的分布特性和结构特征，刻画数据中的多变量关系和相互作用。然后从该模型中抽样或生成合成数据。如果模型能很好地代表真实数据，那么合成数据将具有与真实数据相似的统计特性。以 ChatGPT 为例，它深入研究了人类写的数十亿例文本，分析了词语之间的关系，并构建了一个模型来理解它们是如何组合在一起的。在生成文本时，每一个单词的选择也都取决于它前一个单词出现的统计概率。第二种生成方法并不来源于真实数据，而是通过使用现有模型或者人类专业背景知识来创建。现有的模型可以是某个过程的统计模型，也可以是模拟模型。模拟可以通过游戏引擎等方法创建，如最近火爆的 Sora 文生视频模型，里面用到了由游戏引擎

（Unity、Unreal Engine 5 等）合成的视频数据作为训练集，以提高生成质量。

根据用于训练的 AI 类型，可以将合成数据分为应用于生成式 AI 和判别式 AI 训练两类。应用于生成式 AI 训练的通常有媒体合成数据，即由模型和算法合成的视频、图像或声音。文本合成数据，即在自然语言处理中由模型生成的文本。而判别式 AI 训练（分类或回归）所需的通常是表格合成数据，类似真实生活中数据记录或表格的合成数据。

4.5 合成数据在模型训练中的作用

基础大模型训练所需的数据类型包含两大类，一是用于预训练的世界知识，二是用于对齐的数据。合成数据作为真实数据的一种替代，现阶段虽然在预训练占比不高，但未来发展潜力巨大，可作为一个“新物种”密切关注；目前合成数据多应用于提升对齐阶段的数据获取效率，增强模型安全和可靠性。

4.5.1 预训练语料的新物种

模型预训练阶段是通过大量无监督学习构建基础能力，掌握世界的规律。大语言模型需要各类世界知识，包括网页、书籍、新闻、代码等；而多模态又需要视频、图片、音频等语料。那么合成数据作为新物种，能对模型的训练语料起到哪些补充作用呢？

首先，合成数据可应用于多模态数据的生成。最近火爆的 Sora 文生视频大模型，里面用到了大量由游戏引擎合成的视频数据作为训练集，以提高生成质量。此外，利用模拟器生成的多模态场景数据还广泛应用于具身智能机器人、自动驾驶、AI for Science 等场景的训练。利用模拟模型生成多模态数据可以更好满足模型对训练数据差异化的需求，例如通过有效“过采样”（随机复制少数样例以增大它们的规模）罕见事件或灾难性事件，以确保模型能够针对更广泛的输入保持鲁棒性。而伴随生成式人工智能走向更通用，模型训练将不仅从文字中学习，也会从声音、图片和视频中学习，就更需要多模态的训练数据。因此，我们判断通过合成的多模态数据进行训练的需求还会持续且大幅增加。

其次，合成数据还可应用于高价值领域知识的生成。核心是合成数据能通过对现有数据的深加工，将之前不能被用于训练的数据转化为可用，提升模型对数据利用的可能性。例如工业制造领域，利用合成数据，可以把生产、制造等工艺流程相关的原始数据，结合行业知识图谱，转化为可供大模型学习的工业语料，以缓解行业语料短缺的问题。该过程分为三步：一是将原始数据（Data）转变为信息（Information）：即将非自然语言描述的内容（如工艺生产中的操作行为或时序数据）转化为大模型可读的结构化信息（操作记录）。二是将信息提炼为知识（Knowledge）：仅有操作记录并不能直接提供有效知识，但将多条结构化信息与行业的知识图谱、专家经验相结

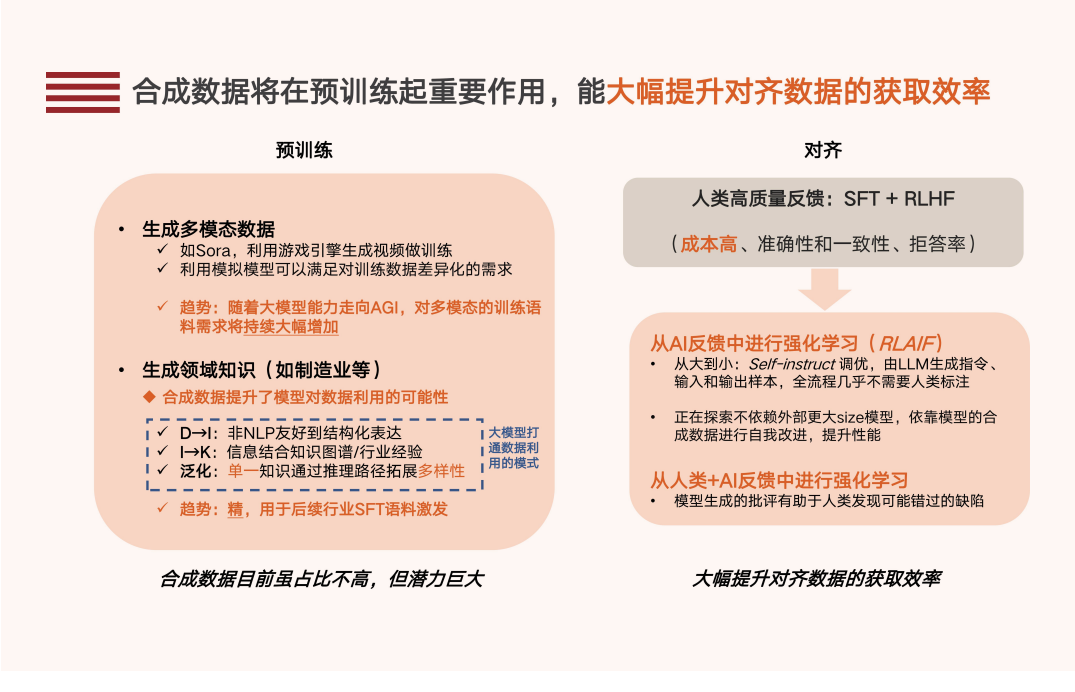
合，可以产出有价值的行业知识（如在什么温度下应该如何操作，好处是什么）。三是将得到的知识泛化：利用大模型的推理能力，将相对单一的知识进行多样性拓展，积累更丰富的行业语料。由此看出，大模型可以利用原始数据、信息、知识等不同层次的内容，打通数据利用的模式。我们判断，通过合成数据拓展对数据利用的可能性，生成领域知识的趋势是“精”，即对语料质量要求高，且是不可或缺的。因为大模型只有在预训练中学习过领域知识，才能在后期利用行业语料进行 SFT 训练时激发出更好的效果，更容易应用于垂直领域。

综上，我们认为合成数据作为预训练语料的新物种，发展潜力巨大，特别是在多模态数据和领域知识生成方面值得密切关注。

4.5.2 提升对齐语料获取效率的加速器

对齐数据以人类高质量反馈为主，包含监督微调阶段和基于人类反馈的强化学习。此方法主要在以下几方面遇到问题：一是数据获取的成本更高，二是人类评估的准确性和一致性，三是模型通常选择避免回答敏感和有争议的问题，降低模型的整体效用。如果引入合成数据作为真实数据的补充和替代，能否缓解这些问题呢？

合成数据最大的优势是可以大幅提升对齐数据的获取效率，“如果掌握了合成数据技术，对齐的成本可能会降低好几个数量级，或用一样的投入产生更大数量级的数据，竞争格局就会发生变化”。这种对合成数据的应用是



“从人工智能反馈中进行强化学习（RLAIF）”。通常是用一个较大规模模型产出合成数据，生成指令及输入和输出样本，过滤掉无效或重复信息，自动化微调出性能较好的小模型，全过程中几乎无需人类标注。这不仅大幅降低了标注成本，也能缓解人工对齐导致模型对敏感问题拒答的情况。例如斯坦福大学发布的 70 亿参数对话大模型 Alpaca，正是采用此类自我指导（Self-instruct）方法，用 OpenAI 的 API 自动生成指令数据进行微调。还有一种基于 RLAIF 新思路探索，希望在不引入外部模型的前提下实现自动化微调。例如自我对局（Self-play），在满足一定条件时，利用合成数据进行自我对抗微调（t+1 代的模型尝试将 t 代模型的输出与真人的输出区分开），得到了比 RLHF 更好的效果。再如 Claude3 用到的宪法式 AI，让 AI 系统在遵循预先设定的原则下，使用模型自身生成的反馈和修正意见来进行自我改进，得到一个既能生成无害内容，又不规避有害问题的模型。同时另一种对合成数据的应用是“从人类和人工智能反馈中进行强化学习（RLHAIF）”，该方法整合了人类和 AI 元素以提供监督。有研究表明，在利用 AI 协助人类评估模型有效性时，模型生成的批评有助于人类发现可能错过的缺陷，提高人类评估的准确性。

4.6 解决训练数据供给不足的新方案

高质量数据是大模型技术发展的主要瓶颈之一，可供大模型学习的数据类型较多，但能够进一步拓展大模型知识边界、推动大模型推理、泛化等关键能力提升的数据更多偏向于视频、图片等多模态数据，以及特定行业中的领域知识数据。此类数据主要来自于人类的创造、制作和经验积累，其规模、类型和质量因客观条件的不同存在较大差异。在大模型强大的无监督数据学习能力面前，大模型的数据需求快速经历了从量到质的转换，能够被大模型更为直接地利用、可以进一步提升大模型关键能力、帮助大模型生成内容更符合人类习惯和要求的高质量数据，成为了最为关键的数据类型。对于提高此类高质量训练数据的供给，现行的主要方案侧重于构建更为开放、包容的高质量数据源，包括建立具有公共或准公共属性的高质量数据集，鼓励行业数据的进一步共享，放宽对于训练数据的权属保护规则等。而合成数据为模型数据供给提供了新的技术方案，将合成数据应用于大模型训练数据中，可以从以下三个方面帮助解决高质量训练数据供给不足的问题。

其一，合成数据解决了部分类型的真实世界数据难以观测的问题，拓展了训练数据的多样性。传统上看，通过生成“边缘情况”（如极端天气、罕见病）或者真实世界中的“潜在隐患”（如金融诈骗等安全风险），可以弥补因为样本分布不均衡导致的客观限制。在输入端纠正数据在采集和处理过程中引入的偏误，提高数据分布的合理性和客观性。面向未来，利用合成数据技术生成的仿真数据（如游戏引擎生成的视频），以及对于大模型难以直接使用数据的加工和改造形成的新型数据（如领域知识），可以提升模型对数据利用的可能性，对于推理、泛化等大模型核心能力的突破将起到更为显著的作用。

其二，合成数据和真实世界的配合使用提高了模型的安全性和可靠性。在 LLM 中，合成数据将更为广泛地应

用于模型对齐阶段，可以提升模型对齐能力，解决基于人类反馈的强化学习过程中人类回答标准不统一，因知识欠缺造成问答准确性不足，以及人类提供反馈成本较高的问题。以高性能模型生成得到的高质量合成数据，以知识蒸馏的方式帮助轻量级模型进一步的监督学习，并为下游开发提供准确、高效的对齐数据来源，从整体上提高各种规模尺寸模型的性能，促进模型安全。在图像领域，合成数据可以弥补对抗样本稀疏的缺陷，将合成图像数据和普通图像数据按照一定比例进行混合，可以提高视觉模型对图片的识别和判断能力，即使在普通数据样本完全缺失的情况下，使用合成数据进行图像识别训练，也可以得到接近普通数据样本训练的效果，从而提升图像识别的鲁棒性。

其三，合成数据可以替代个人特征数据，有助于用户隐私保护，解决数据获取合规性的问题。例如，当合成数据用于推荐系统，可以降低后者对个人信息的依赖。传统的直接利用个人行为特征数据进行推荐，模型并不能从文义角度理解用户的需求，为了提升“猜你喜欢”的准确度则需要获取和分析大量的用户行为特征信息。在推荐系统等涉及个人隐私信息的判别式模型中，通过与大模型的结合可以有效缓解该问题。首先，利用生成器自动产出个性化提示词（即合成数据）用于模型优化；然后，发挥大模型对文义的推理能力，可以更好地预测用户的实际需求。用户和大模型进行简单沟通后，由大模型代为执行推荐，在提升推荐匹配度的同时还可以降低推荐模型对个人特征数据的依赖。推荐模型不再高度依赖个人特征信息，也为隐私增强技术的加入提供了操作空间，在合成数据的生成过程可以加入差分隐私等去标识技术，推荐系统在不识别特定用户的情况下也能良好判断用户的实际需求，进行针对用户实际需求而非臆测性、推断性的推荐。

4.7 在发展中治理的合成数据

其一，相比于对合成数据量的扩增，在应用中要更重视质的提升。首先，在语料中使用占比更高的仍然是来自真实世界的数据集，合成数据未被用于大规模替代真实数据进行预训练。相反，如果此阶段过多引入合成数据，可能会影响训练数据分布，从而导致模型对世界知识的理解产生偏差。其次，合成数据的总体规模也会受到模型生成能力和生成速度的限制（例如按照当前的合成图像数据生成速度，在 A100 GPU 上每个图像生成时间大约为 0.8s；启用 xformer 时，在 V100 GPU 上每个图像的生成时间约为 2 秒）。因此，更重要的是关注生成合成数据对客观世界模拟的准确性，更好满足模型对训练数据差异化的需求，以及拓展模型对训练数据利用的可能性。较为通用的方案是按照一定比例将合成数据与真实世界的数据进行混合，用于模型优化，提升模型准确性、鲁棒性和安全性。

其二，合成数据本身具备良好的安全性，在后续使用中较为可靠。用于模型优化训练的合成数据目的在于替代普通优化数据提高模型的对齐能力和垂类应用效果，要达到此目的，合成数据安全性和真实性不低于真实世界的数据，否则使用合成数据并不能更好地提升模型性能 —— 如果合成数据的质量低于真实数据的数据，则可能造成模型性能不升反降，使用合成数据的价值也将大打折扣。现实情况来看，合成数据往往也是通过高性能模型生成而得来的，此类模型具有良好的安全防护机制，能够有效控制生成内容的安全性，因此产生的合成数据在下游利用中可靠性良好，不会带来“数据 - 模型自我循环”导致的模型失控问题。

其三，对合成数据仍需设置相应的安全管控策略，确保模型整体的安全性不会因为合成数据的使用而受到影响。一是加强对合成数据质量的评估检测。合成数据和其他类型的训练数据一样，需要不断提高准确性和可靠性，而为了保证合成数据具有可用性价值，其准确性和可靠性要高于普通的真实世界数据。二是为合成数据设置备用数据集。合成数据在模型训练中的使用还处于探索阶段，需要更为审慎地观察不同类型、模态和配比合成数据对模型性能带来的影响，并为合成数据准备备用的真实世界数据集，当模型能力和安全性评测、红队测试等监控指标出现异常时，及时介入并采用备份的数据集继续模型训练和应用，保证模型的稳定性。三是建议对用于模型优化、对齐的合成数据在适当环节引入人类参与。例如，对用于对齐阶段生成的问答对和其他媒体格式内容，在进行模型优化前进行人工抽检，确保后续模型调优和对齐的质量。

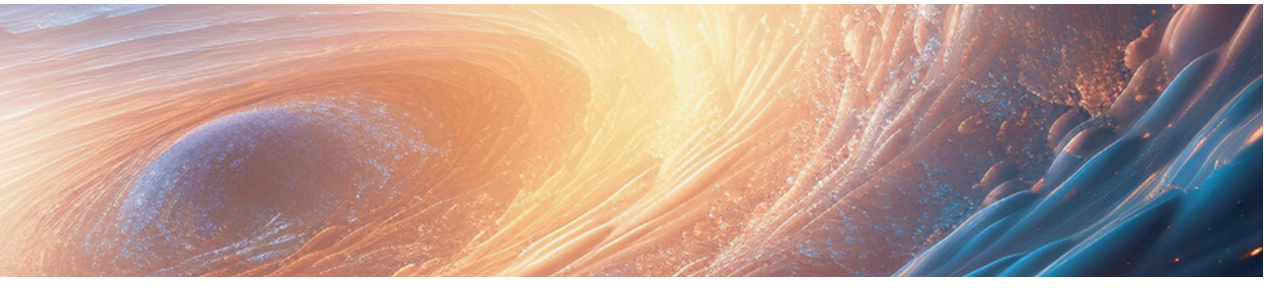
05

对大模型训练数据治理的思考

5.1 大模型对训练数据的使用特点

首先，在个人信息方面，模型训练阶段不依赖个人信息，对公开个信的使用属于合理使用。人工智能技术从依赖个人信息的决策模型转向以大模型为代表的生成式 AI，反映出数据需求的深刻变革。具体而言，大模型的技术核心在于模拟人类思维进行内容创造，输入端的训练数据侧重全球知识和高质量语料，而非个人信息，即便在前端降低个人信息在训练数据中的含量和真实性，均不会对模型最后所展现的性能产生较大影响。其次，即便大模型训练语料中涵盖个人信息，大模型研发者已按照相关安全要求，采取技术手段进行数据清洗、去标识化、匿名化等操作，对其中所涵盖的个人信息进行了最大化的去除。而剩余的通过爬虫等技术获取的位于公共领域的个人数据，大模型对于此部分数据的使用应构成合理使用的范畴。

其次，大模型对版权类训练语料的使用是转换性使用，属于合理使用或法定许可。大模型对于版权作品的使用，并不是以欣赏作品原有价值为目的而进行利用，或对原有作品内容进行复制和传播从而替代原有作品，而是为了掌握客观规律并培养模型的基础能力，就如给人类进行教育需要对其进行广泛的知识授予一般。有鉴于此，用版权类数据对模型进行训练，不应被视为“复制式拷贝”的版权侵权行为，而应属于转换性使用的范畴，并应构成“合理使用”或“法定许可”。目前，已有法律实践在模型训练使用版权作品方面做出突破，如欧盟《单一数字市场版权指令》为符合条件的“文本和数据挖掘”设置了豁免例外，日本对《著作权法》的修订将“不以欣赏作品原有价值为目的”的大模型数据训练纳入到合理使用的范畴等。



此外，模型训练已经尽可能地采取了相关合规方案，来减少生成式人工智能造成知识产权侵权的风险，具体包括：（1）从真实权利人处购买具有知识产权权利的数据库；（2）使用有合法授权的开源数据集；（3）避免跨越技术措施的爬取。

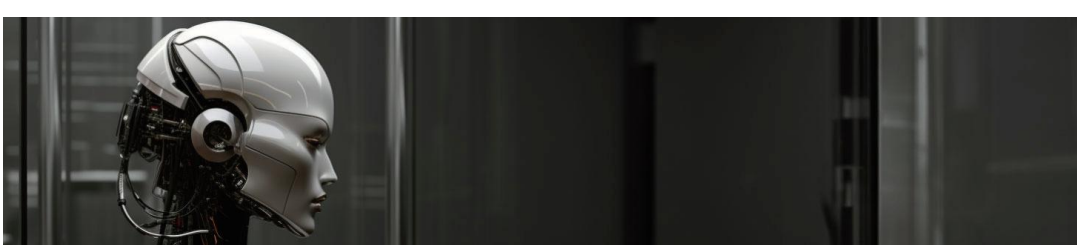
5.2 大模型训练数据合规的治理之智

基于大模型对训练数据的使用特点，应构建顺应模型发展的新时代的数据治理制度。

一是重视数据的可及性，从输入端的前置使用限制，替换为输出端的管控和事后救济。《生成式人工智能服务管理暂行办法》从 2023 年 4 月征求意见至 7 月正式公布期间，充分考虑了我国大模型发展的实际需要，在训练数据合规方面也适当放宽了要求，如删去训练数据“不含有侵犯知识产权”的表述，调整为在训练过程中“不得侵害他人依法享有的知识产权”。我们看到目前新制度的创新尝试和旧有制度的延续使用，仍在大模型训练前端的数据获取和使用方面对模型研发者施加了较为严苛的前置性合规要求，如训练数据需记录所有个人信息并取得个人知情同意，以及识别训练数据中的知识产权侵权风险语料并进行删除等。此外，训练数据的供给不足一定程度上限制了我国大模型尤其是基础模型的赶超式发展，对于训练数据的使用管住输出端的事后风险、放宽输入端的事前限制是务实的政策选择。

二是提升模型安全训练数据的供给，鼓励安全类数据集的开放共享。大模型的能力和表现非常依赖于前端数据的输入，而安全数据集作为高质量模型训练数据之一，具有正外部性，将有助于大模型的人类价值对齐，并对什么是安全和不安全的内容，以及怎样正向地回答这些问题进行系统性的了解和学习。因此，覆盖全类别、横跨多领域的安全数据集的开放共享，将显著提升人类价值观对齐在性别、职业、种族、无障碍领域，并有助于提升大模型后端内容生成和输出的无毒性、安全性和可靠性，帮助大模型更加得体和正面地应对更广泛的问题。

三是应用新技术以提升训练数据的合规性和安全性，比如合成数据的使用可以增强对个信的保护。一方面合成数据的应用可以减少对具有可识别性的个人特征数据的依赖，另一方面合成数据通过差分隐私“加噪声”的技术，可以有效实现去标识化，从而有助于增强对个信的保护，更好解决数据获取合规性的问题。



06

政府与社会力量协同的训练数据生态

本节从政府和社会力量两方面讨论大模型训练数据的来源。通过中美对比的现状，分析两者的差异性，以及对我国人工智能领域数据要素发展的借鉴作用。

从政府视角看，哪些公共数据可以支持大模型训练？我们梳理了以下几种：一是，经过权威认证或凝聚共识的知识，如专利文档、上市公司财报、法院判例 / 裁判文书、医疗诊断记录、政策文本等，除了可用于预训练语料，还可应用于行业大模型监督微调或外挂语料库建设。二是，具有科研属性的数据，主要特征有长周期、大规模、多模态、来源清晰、描述详细、可使用，如天气、医疗、地球科学、基础科学领域等，用于 AI for Science，让模型提升从复杂数据中提炼规律，提升精准预测的能力，同时拓展 AI 大模型在更多领域中应用。三是，科研期刊论文，用于提升模型上下文的理解能力和逻辑推理能力。

而社会力量整合政府开放数据与网络公开数据，在拓展广度的同时，提升精细度和专业性。“广”的层面，社会力量将公共数据与网络公开数据融合后做进一步清洗和加工，形成具有多样性、大规模、高质量特点的预训练数据集。此外，社会力量还可以通过合成数据等技术手段，拓展模型对数据使用的可能性。“齐”的层面，社会力量通过大量高质量反馈做数据标注，将模型产出与人类价值观对齐。“专”的层面，也会整合领域知识和经验，促进语料的流通和共享，提供行业大模型所需的高质量、专业性的数据供给。由此可见社会力量在大模型训练语料中所起到的主导作用。

那么，美国与中国在获取大模型数据方面的做法有哪些不同呢？

6.1 美国的现状

美国联邦政府在公共数据中承担了“应开尽开”的职责，由社会力量来探索数据的应用。政府开发了专门针对 AI 训练数据的开放平台，并针对公共数据和科研数据进行质量维护和运营管理，在保证数据可用性的同时降低公众使用门槛。公共数据开放的范围限定在政府数据，包括各级政府及政府资助的大学和研究机构。

在开放共享阶段，联邦政府会对与 AI 相关的数据做标识、在数据量大时做“上云”处理、定期更新、分类、清洗、标注、结构化、并确定分级开放权限。在开发利用阶段，政府会提供便捷的用户检索服务、提供数据接口

（API）。在科研论文方面，设立 PubMed 论文检索系统，整合国家医学图书馆下属的 3 个论文数据库资源，记录了 3600 万 + 条生物医学文献的引用和摘要，并提供原文链接。在科研属性公共数据方面，国家气象和海洋局（NOAA）从卫星、雷达、船舶等来源每天新产生数十 TB 数据，按季度更新 150 个数据集，因数据量庞大存储在云端。为方便公众开发利用，提供了数据集 API 接口。在权威认证的知识方面，如法院的裁判文书是很好的结构化数据，对于训练法律大模型价值很高。美国遵循“公开是原则，不公开是例外”的理念，除了隐去涉及国家秘密和个人隐私的信息，联邦和地方法院都实现了公开，并提供了 API 接口供调用。在医疗领域，含有医 - 患 - 药信息的诊疗记录、CT 图片及结果标注构成的医学影像数据、基因组与疾病筛查数据等对于医疗大模型训练有较高价值，以国立卫生研究院（NIH）为主的机构在确保隐私保护的前提下对公众实现分级分类开放（139 个医疗健康数据库，包含 9 个医疗影像数据库，拥有超过 30 万张 CT 图像及标注对、20 个基因组数据库），供社会力量使用。

美国的社会力量整合政府的开放数据与网络的公开数据，提升数据精细度和专业性，形成以开源为主的高质量训练语料。社会力量主要有开源 / 非盈利组织、互联网公司研究部门、学界、多类型机构合作组成。数据集以开源为主，站在前人的肩膀上不断迭代。以在大模型中被广泛应用的，由开源组织 Eleuther AI 开发的 825GB 高质量英文语料库 The Pile 为例，在 22 个子数据集中，来源于政府公共数据的有 4 个（PubMed 数据库、商标专利数据库、卫生研究院数据等），这也体现了语料中不同类型数据有效搭配的重要性。在行业大模型中，社会力量对领域数据集的专业性也起到了重要贡献。以把大模型当做大脑来辅助运行的具身智能机器人为例，Google DeepMind

美国社会力量参与大模型预训练数据集：以The Pile为例

构成部分	原始数据大小 (GB)	是否政府提供	数据类型	
Pile-CC	227		网络	网页爬取
PubMed Central	90	是	科研	生物医学文章
Books3	101		书籍	
OpenWebText2	63		网络	Reddit论坛
ArXiv	56		科研	论文数据库
Github	95		其他	代码数据库
FreeLaw	51		科研	法院意见
Stack Exchange	32		网站	问答网站
USPTO Backgrounds	23	是	公共数据	专利数据库
PubMed Abstracts	19	是	科研	生物医学文章摘要
Gutenberg	11		书籍	
OpenSubtitles	13		对话	电影/电视英文字幕
Wikipedia (en)	6		网络	维基百科
DM Mathematics	8		其他	数学问题集合
Ubuntu IRC	6		对话	
BookCorpus2	6		书籍	
EuroParl	5		对话	多语言语料库
HackerNews	4		对话	
Youtube Subtitles	4		对话	Youtube 字幕
PhilPapers	2		科研	哲学论文
NIH ExPorter	2	是	公共数据	国立卫生院经费数据
Enron Emails	1		其他	安然公司邮件数据集
The Pile	**825**			

资料整理自：The Pile: An 800GB Dataset of Diverse Text for Language Modeling

- 社会力量整合政府开放的科研/公共数据、与其他公开网络数据，形成高质量语料的典型案例
- 美国的社会力量主要有**开源/非盈利组织、互联网公司研究部门、学界以及多类型机构合作**
- The Pile 是由美国**开源组织*Eleuther AI*** 开发的825 GB英语文本语料库，被广泛利用于大语言模型训练
- **来源多样**的22个训练子集可以提升模型**跨领域知识和下游任务的泛化能力**
- 社会力量倾向于**数据集的开源**，可以站在前人的肩膀上不断迭代，避免重复造轮子

联合 33 家学术实验室，汇集了来自 22 种不同机器人类型数据，涵盖 100 多万条片段，展示机器人在 15 万项任务上的表现，创建 Open X-Embodiment 开源数据集。基于该数据集训练的具身智能模型，解决了机器人在特定任务专业而通用能力差的难题，成功率提高 50%，技能表现提高 2 倍。此外，在合成数据领域，美国的发展也显示出积极的趋势和广泛的应用前景，比如微软在其投资组合中就包含了诸如 hazy、Unstructured-IO 等合成数据公司。

在政府与社会力量协同的方面，美国联邦政府发挥了 AI 训练数据“汇聚融合”的角色。为巩固美国在 AI 领域的竞争优势，由政府主导推动为期 6 年的国家人工智能研究资源 NAIRR 计划，让 AI 研究者获得更多算力和数据资源。计划的原则是尊重社会力量的专业性，作为经营主体的指导委员会中有多位来自 AI 业界和学界的资深人士。NAIRR 在数据资源整合中发挥的作用体现在，联邦政府通过建立数据资源服务平台，汇聚政府与社会力量的开源数据资源。通过建立统一的数据汇聚标准，规范数据描述格式，促进多方数据融合。倡导 AI-Friendly 的数据兼容性，将数据集整理和格式化成易于 AI 算法处理和学习的形式，如文档的电子化程度、版面编排以及相关数据来源的完整性。同时推动多方协作的数据资源开发利用，如运营数据集社区、提供数据搜索服务等。

6.2 中国的现状

我国的公共数据采用主体性质界分，包含各级行政机关在履行公共管理职能中获取的数据，覆盖范围比美国更广，但在开放共享和开发利用程度上仍有不足。如天气数据的开放，在中国气象数据网查询地面逐小时观测资料时，个人用户需注册，且可选范围被限定在 7 天以内；而对比 NOAA，无需注册即可下载，且以地表温度为例，数据最早可追溯到 1951 年。在开发利用中，我国也仅对个别数据集提供了 API 接口。再如法律领域，最高人民法院设立了裁判文书网，除例外情况外统一公布各级人民法院的生效判决书。但近年公开的数量有明显下降趋势，2020 年上网文书 2300 多万，而 2023 年截至 12 月仅公开 300 万。另 2024 年 1 月将启用“全国法院裁判文书库”，仅法院人士在内网可查询。在医疗领域，对于模型训练价值较高的医疗影像、基因组数据开放程度非常有限，社会力量的探索呈现“散点状”。

我国的社会力量主要是结合海外优质开源数据集及中文语料，产出训练数据集。以阿里巴巴的“通义千问”大模型为例，训练数据来自公开来源的混合数据，以中文和英文为主。而中文语料主要来自知乎、百度百科、百度知道等公开网络数据，来源于政府的公共数据非常少。从总体看，中文语料库的开源情况不如英文普遍，据 AI 应用开放社区 Hugging Face 数据统计，中文开源数据集数量仅占英文开源的 11%。在行业大模型中，社会力量对行业数据集专业性有一定贡献，推动了在交通、政务、医疗等领域的应用。整体看，用领域知识训练大模型仍面临困难，第一是领域知识积累的专业门槛高、时间周期长。第二是企业出于商业利益和知识产权考虑，对领域知识共享意愿度低。第三是因为我国公共数据开放不足，导致部分行业缺少优质的数据供给。在这种情况下，如果还要试图缩小已经开放的公共数据范围，那么高质量语料短缺的问题将更为突显。

我国尚未形成对大模型提供有效供给的数据资源生态。相比美国政府以公共数据开放服务于训练语料，社会力量以融合公共数据和网络公开数据提升语料广度、精细度和专业性的生态模式，我国可供大模型训练的有效数据资源呈现碎片化分散状态。中文语料、科研成果等高质量数据集开放程度低，企业用于训练的语料来源不清晰、权属不明确，开源后存在一定的合规隐患，使得企业更倾向于自采、自用，大模型数据流通机制尚未形成。此外，由于过多依赖删除手段治理，导致网络上有中式价值观的高质量公开语料供给较少。

训练数据从何而来：美国和中国的对比

训练数据来源	美国	中国
政府	• **联邦政府科研数据&公共数据“应开尽开”及运营维护** • 促进联邦政府数据+业界数据“汇聚融合”	• 公共&科研数据开放程度不高 • **部分地方政府有所尝试，**如： ✓ *北京：高质量语料库撮合交易* ✓ *深圳：探索市级公共数据开放和运营*
社会力量	• **基于政府开放数据及海量网络数据加工处理形成预训练数据集** • 贡献以**开源为主**，站在前人肩膀迭代优化 • 标注人员对模型产出提供高质量反馈	• **结合海外优质开源数据集及中文语料拓展产出中文训练数据集** ✓ *训练数据中来自政府的语料很少* ✓ *中文语料库开源情况不普遍* [1)] • 标注人员对模型产出提供高质量反馈

注1）：据AI应用开放社区Hugging Face数据统计，中文开源数据集数量仅占比英文开源的11%

07

阿里巴巴集团在大模型训练与应用的探索

以上阐述了大模型训练数据的技术原理，本节以阿里巴巴集团在大模型训练和应用中的部分案例，简要说明训练数据在产业中的实现路径。在大模型训练数据的处理和应用中，阿里巴巴集团整合优质的中文语料与海外开源数据集，在确保数据合规性的同时不断迭代，优化训练数据质量。

在探索不同数据类型之间的有效搭配时，阿里巴巴达摩院在语料学习顺序中进行了“数据课程”的设计，在预训练和监督微调阶段之间，引入了“持续预训练”环节。因为在达摩院推出面向东南亚语大模型 SeaLLM 时，面临着东南亚区域语料供给稀缺的问题。遵循数据课程的逻辑，达摩院基于 Llama-2 模型，在预训练第一阶段使用语言识别工具，只保留英、中、泰、越南、印尼语言的文档；而在预训练第二阶段筛选高棉语、老挝语、马来语、缅甸语等特定语料专项学习，通过持续预训练来扩展词汇量，专门针对东南亚语言进行优化，以确保模型能够学习到丰富的语言特征和文化背景。

在涉及个人信息内容时，由于大模型训练不依赖个人信息，因此在训练阶段会主动采用技术手段从源头减少个人信息收集、降低个人信息在训练中的比例和真实性。在实践中，由于预训练阶段语料数量巨大，常采取“关键词 + 正则表达式匹配”的方式检测个人信息，然后执行删除或者模糊化操作。

在合成数据的探索和应用中，在电商场景尝试通过合成数据实现 LLM 与推荐系统结合，更好地推理用户真实需求。其技术路径可概括为三步：第一，通过提示词生成器，由算法根据上下文信息、用户需求信息构建定制化的提示词；第二，用合成数据对预训练大模型做优化，这通常用效率较高的“提示词微调”方法实现。第三，利用微调后大模型的推理能力，给出更为合理的、用户能够理解的推荐理由，让用户更容易理解推荐内容。引入 LLM 之后的推荐系统可以在效能提升和隐私保护两方面具备优势。从效能提升视角看，一是推荐商品的丰富性提升，由于大模型具有推理能力，通过推荐思路可以无限向外扩展，丰富的内容可以引导用户的发现性。二是无需冷启动，由于大模型具有少样本甚至零样本学习的能力，推荐可以依赖大模型内在的客观世界知识，不需要大量场景数据的积累，就能快速迁移和复用。三是可解释性增强，将大模型的推荐思路以推荐理由的形式外化给消费者，可以让用户更好理解推荐的逻辑。甚至可以通过 LLM 与用户的多轮交互，响应实时诉求，做到可交互性。从隐私保护视角看，引入合成数据会降低推荐系统对用户行为等数据的依赖程度。此外，用差分隐私的语言模型可以创造一批“加噪声”的合成数据，这些数据在统计上代表了原始数据，但不包含任何个人可识别的信息。用这些合成数据来训练，即使模型学到了很多东西，也不会侵犯到真实用户的隐私。

08

以更开放和务实的方式解决高质量训练数据供给

促进我国人工智能数据体系建设，需要理解大模型对训练数据的实际需求，数据质量与模型能力的关系，综合利用政府和社会力量各方资源推动数据的开放、开发和利用，构建共享、共创、共赢的合作生态。

在意识层面，制度设计要给技术发展预留空间。正如国务院研究室副主任陈昌盛在“关于当前促进数字经济发展的六个优先”中提到的，“数据的可及性优先于数据的确权”。随着模型能力提升和模态扩展，高质量数据类型的演进具有不确定性，难以预判，因此在不违反国家安全、个信保护、企业商秘三条红线的前提下，对大模型训练数据的使用应持更开放的态度，不要过多在输入端做管控，要给技术发展预留空间。而对待剩余风险，可以更多采用输出端限制和事后救济补偿的原则。因为在技术原理上，一方面大模型训练不依赖个人信息，另一方面对版权类数据的学习属于转换性使用，并非直接的拷贝和复制，可被视为合理使用。另外，对正在发展中的技术，应以促进开发利用为目标确定保护规则，推动模型能力建设，特别是对作为中间产品类型的合成数据，不宜过早过度保护。

在操作层面，高质量数据要素的供给离不开政府与社会力量的市场化分工协同。如同人工智能的发展历程一样，如何构建高质量数据也并没有标准答案和成功先例。此类问题要想取得突破，正如著名经济学家许成钢所说，需要的不是政府直接干预，而是大量的自由探索和大批的独立研究。特别是在具有专业性和需要试错迭代的领域，基于市场优胜劣汰可以更高效的判断好坏、配置资源。在政府侧，对可用于模型训练的公共数据鼓励“应开尽开”，在数据开放过程中不要过多预设使用场景。在社会力量侧，企业和相关机构“应试尽试”，通过在数据混合与数据课程环节的不断迭代，寻找发挥最大价值的“配方”。在该过程中，社会力量本质上是凭借各自对技术和市场的理解，投入时间、人力和算力，探索数据集的构建方法。在市场机制层面，高质量语料效果会在模型训练和应用中得到检验，其价值可依据商业合同对价按效果付费，而不是按资源占用规模。

对于有确定性、已经研究清楚的数据类型，要坚决促进利用。对有助于模型提升对价值观引导能力的中式价值观语料，以及增强对物理世界专业性理解的科研数据，应高度重视开放共享和开发利用，涉及到版权类语料要旗帜鲜明地扫清制度障碍。特别是对于受财政支持的科研或文化单位所有的知识产权类价值观语料，应尽快向社会公开用于基础大模型训练，同时可基于非营利性成本补偿原则明确合理收费标准，如媒体的主流价值观数据，国家图书馆电子化图书、历史典籍、数字报纸、科研期刊和论文等。

2024 年 5 月第一次印刷

完整电子版下载，请扫码关注阿里研究院公众号获取

AliResearch
阿里研究院

阿里研究院是阿里巴巴集团的内设智库机构，多年来扎根于阿里巴巴丰富的数字科技商业生态，依托前沿的产业实践和大量的创新案例，围绕集团“用户为先，AI 驱动”的战略重心，聚焦于科技创新、数据和算法治理等领域的研究。

阿里研究院数据经济研究中心

聚焦于数据流通、数据跨境、数据安全、模型训练数据、数据资产等数据政策研究。在产业数字化方面，关注数据要素对于传统行业的赋能，促进数据的合规流通与高效利用，推动中小企业数字化转型。在数字产业化方面：关注数据要素与人工智能结合的价值释放，在大模型数据领域提供科技企业的洞察和建议。

阿里研究院 AI 治理研究中心

依托阿里巴巴集团先进的 AI 科技能力、丰富的应用场景与负责任的技术生态，聚焦于 A 风险、AI 大模型安全、AI 大模型开闭源生态、AI 大模型出海等 AI 治理政策研究，基于技术理性的风险观，关注大模型发展与安全平衡与取舍，为政策制定实施提供科技企业的智识建议。

阿里研究院 AI 产业研究中心

依托于阿黑巴巴集团在人工智能领域的全面布局，长期追踪人工智能芯片与模型的技术推进，以及基础产业生态和应用的实践落地。在算基础设施的规划与布局、能耗优化、大模型能力发展、大模型评测体系、产业应用案例深入研究、新技术与应用趋势前瞻探索等方面，开展系统性与深入性的研究工作，旨在为政策制定与实施提供基于产业深度洞察的策略建议，以促进 A 技术的健康发展和产业的创新升级。