专题 大模型训练数据

构筑 AI 时代的基石——科学数据

洪学海

1. 中国科学院计算技术研究所，北京 100190

*Email: hxh@ict.ac.cn

人工智能（AI）技术以前所未有的深度和广度重塑着科学研究的范式。高质量科学数据正日益凸显其作为“新型基础设施”与“战略资源”的基石性地位。《中国科学数据》大模型训练数据专题汇聚的多个数据集成果，指向一个核心命题：在 AI 时代，数据是新型科研范式的基础设施，更是驱动关键领域突破的战略资源。

科学数据的深度与广度，直接决定了 AI 模型的认知边界与决策精度。本专题成果覆盖领域极具代表性：从关乎国计民生的洪涝灾害精准监测，到支撑民族文化传承与信息化的藏语语音识别、文本摘要；从赋能智慧服务的旅游知识图谱、汽车维护知识库，到服务于社会治理与国家安全的法律问答、外交辞令分析；再到探索人工智能核心能力的复杂推理、价值观对齐与文本检测。这种广泛的覆盖性表明，数据驱动的智能技术正深度融入自然科学、工程技术、社会科学及人文领域的方方面面，成为跨学科创新的“通用语言”和共同基础。

与此同时，数据的构建呈现出显著的深度聚焦特点。研究不再满足于数据的简单汇集，而是针对特定领域的核心挑战进行“靶向”设计。例如，为解决低资源语言处理难题，构建了专业的藏语安多方言、拉萨话语音识别数据集；为提升大模型在专业领域的复杂推理与对话能力，构建了涵盖医疗、法律与金融的多轮专业咨询对话数据集；为直接增强大模型的逻辑推理与谬误识别能力，专门设计了数学难题数据集与逻辑谬误数据集。这种深度聚焦意味着数据工作已从支撑性、辅助性角色，转变为直接参与定义科学问题、攻克技术瓶颈的前沿研究活动本身。

科学数据的价值，不仅在于其作为模型“燃料”的基础作用，更在于其推动科学问题发现、研究范式革新的深层动力。它使得科研人员能够从海量、多元的信息中抽丝剥茧，构建起对复杂系统的深度认知。在知识密集型领域，如法律、医疗、金融，结构化、多轮对话的专业咨询数据集，为训练大语言模型理解复杂逻辑、进行专业推理提供了高质量的“教材”，使其能够模拟专家思维，提供更具可解释性的服务。面向法律问答的微调数据集，通过覆盖高频诉求场景的真实咨询记录，有效提升了模型在法律意图理解、法条引用及逻辑推理方面的能力，推动了法律智能服务的落地。更进一步，专门针对模型能力短板构建的数据集，如面向复杂推理的数学难题数据集、逻辑谬误数据集，直接瞄准模型在深层推理、谬误识别方面的弱点进行“靶向训练”，旨在从根源上提高模型的思维链质量和逻辑严谨性。这些努力表明，科学数据的构建已从被动的资源收集，转向主动设计以解决特定科学问题、牵引模型能力向更高阶跃迁的创造性活动。

科学数据的构建过程本身，就是一项严谨的科学研究，蕴含着对数据质量、代表性、价值的深刻考量。本专题多个数据集的研究都强调了严格的数据质量控制流程，包括人工审核、多模型评估、去重去噪、隐私脱敏等环节。例如，在价值观对齐数据集的构建中，通过 IFD（Instruction Following Difficulty，指令遵循难度）筛选优先保留对模型能力提升有帮助

的复杂样本，再通过人工审核确保数据符合社会伦理；在多类型推理问答数据集中，采用多模型评分机制从多个维度进行量化评估；在生成式藏文文本摘要数据集中，结合主客观方法对文本与摘要的匹配度进行全面评估。这些质量控制措施，是确保数据可靠性、科学性和可用性的生命线，也是数据工作能够真正赋能科研、服务社会的根本保障。在方法论上，数据构建的创新性同样突出。研究者们综合运用了多种前沿技术路径：从利用无人机获取高分辨率实地观测数据，到基于网络公开资源进行采集、翻译与优化；从采用提示增强、数据演化等策略进行合成与扩增，到借鉴知识图谱等技术对多源异构信息进行结构化整合。这些方法创新不仅提升了数据构建的效率与规模，更拓展了数据内容的多样性与复杂性，为训练更强大、更稳健的 AI 模型提供了可能。

尤其值得关注的是，科学数据在促进学科交叉、服务国家战略与社会需求方面展现出巨大潜力。从赋能智慧旅游的甘肃旅游知识图谱，到支撑地质调查与科普的深成侵入岩类图像数据集，再到服务于外交话语分析的中文外交辞令数据集，数据驱动的智能技术正在与国民经济、文化生活、社会治理的方方面面深度融合。这些数据集不仅为相关领域的学术研究提供了宝贵资源，更体现了科学研究面向国家重大需求、面向经济社会发展主战场的责任担当。

综上所述，本专题所展示的丰富成果，是中国科研人员在科学数据这一基础战略领域辛勤耕耘的缩影和进行的系统性、创新性实践。它们生动诠释了“数据即资产，数据即竞争力”的时代内涵。展望未来，唯有夯实数据这一“智能时代的基石”，我们才能在全球科技竞争中牢牢把握创新驱动发展的主动权。我们呼吁，应从国家科技战略的高度，进一步加强对关键领域科学数据建设的系统性规划与持续投入，建立鼓励数据开放共享的激励机制与标准规范，构建安全、可信、高效流通的数据要素市场生态。鼓励更多的科研力量投入到这一基础性、战略性的工作中来，以更高的标准、更广阔的视野、更紧密的协同，共同构筑坚实、丰富、开放的国家科学数据战略资源体系，为抢占人工智能发展制高点、实现高水平科技自立自强奠定牢固的基石。

论文引用格式

洪学海. 构筑 AI 时代的基石——科学数据[J/OL]. 中国科学数据, 2026, 11(1). (2026-03-27). DOI: 10.11922/11-6035.csd.2026.0031.zh.