人工智能驱动的数据标注：从辅助者到主导者的范式革命

何源
新南威尔士大学，澳大利亚 悉尼 NSW2000
DOI:10.61369/ETI.2025100037

摘　　要： 数据标注作为人工智能模型训练的起点，其效率和质量的高低直接决定了算法能力上限，纯人工标注方式存在代价高、难规模化、质量不稳定等痛点，成为制约 AI发展的“标注瓶颈”。论文将系统介绍人工智能赋能数据标注的发展历程，并提出一个核心命题：AI在数据标注领域的应用从“辅助”正在演变为“主角”。首先分析传统标注方式中成本、质量和规模三难权衡的困境；其次分析以主动学习（Active Learning）为代表的 AI助标注范式，分析其效果与“主动学习悖论”的固有困境；然后重点分析以 LLM和视觉基础模型为代表的新范式，分析其如何通过零样本或少样本等能力重塑标注工作流，以及存在“泛化－特化鸿沟”等新问题；最后分析新范式下的算法偏见、模型可靠性等问题，并展望人机协作的新范式。我们认为，未来数据标注的核心是构建一个人机协同的强大生态系统，在一个大的系统中由基础模型做规模化工作，人类专家进行更高水平的监管、治理和对齐。

关 键 词： 数据标注；人工智能；主动学习；基础模型；大型语模型；人机协同

Ai-Driven Data Annotation: A Paradigm Revolution from Assistant to Leader

He Yuan
The University of New South Wales, Sydney, Australia NSW2000

Abstract: As the starting point of training artificial intelligence models, the efficiency and quality of data annotation directly determine the upper limit of algorithm capabilities. Pure manual annotation methods have pain points such as high cost, difficulty in scaling up, and unstable quality, which have become the "annotation bottleneck" restricting the development of AI. This paper will systematically introduce the development process of artificial intelligence empowering data annotation and put forward a core proposition: the application of AI in the field of data annotation is evolving from "auxiliary" to "leading role". First, analyze the predicament of balancing cost, quality and scale in traditional annotation methods; Secondly, analyze the AI-assisted annotation paradigm represented by Active Learning, and analyze its effect and the inherent predicament of the "paradox of active learning". Then, focus on analyzing the new paradigms represented by LLM and visual-based models, and examine how they reshape the annotation workflow through capabilities such as zero-shot or few-shot, as well as the existence of new issues such as the "generation-specialization gap"; Finally, issues such as algorithmic bias and model reliability under the new paradigm are analyzed, and the new paradigm of human-machine collaboration is prospected. We believe that the core of future data annotation lies in building a powerful ecosystem of human-machine collaboration. In a large system, the basic model will perform large-scale work, while human experts will carry out higher-level supervision, governance and alignment.

Keywords: data annotation; artificial intelligence; active learning; basic model; large-scale language model; human-machine collaboration

引言

（一）研究背景：数据是人工智能的基石

人工智能表现的好坏很大程度上取决于训练数据的质量和数量，这是行业共识。从某种角度来看，高质量标注的数据集就是原始数据和机器可理解知识的融合，亦是现代 AI技术不断突破的源泉[1]。数据标注（标注图像、文本、音频等原始数据）几乎是所有基于监督

作者简介：何源（1997.12—），男，汉族，安徽合肥人，硕士研究生，研究方向：人工智能。

学习的算法得以发挥效用的前提。数据标注产业作为 AI基础产业，其发展深度影响着中国“人工智能 +”行动的深度，尤其是在计算机视觉（Computer Vision）、自然语言处理（Nature Language Processing）等复杂任务上，标注数据的质量决定了模型的性能极限[2]。因此，数据标注不仅是一项技术工作，同时也是 AI产业链的基础设施。

（二）“标注瓶颈”对 AI发展的制约

尽管数据标注不可或缺，但“劳动力密集型”的传统标签却逐渐成为 AI 产业发展的瓶颈。“标注瓶颈”主要体现在成本、效率、质量三大元素间必然存在的矛盾[3]。首先，人工标注的成本居高不下，包括人工成本、人工工具成本、质量返工成本、批量团队管理与数据合规成本[4]。其次，人工标注的可扩展性极差，当数据量达到一定规模，项目管理复杂性成倍增加，知识交流、人员教育成本增高，拖慢 AI 项目的研发进度[5]。最后，标注质量无法保证[6]。标注员的主观偏好、认知习惯乃至文化差异均会导致标注不统一甚至是错误标注，带来模型性能不佳或产生模型的公平性偏见[7]。这些困难构成了一个有组织的“标记税”(Marked Tax)，即每个项目产生价值之前必须支付的一笔相当可观的时间和资本、精力和人力的花销。

为解决这些问题，我国相关部门给出指导意见，要实现数据标注高质量发展，要依托创新，依托标准，解决好产业发展的痛点、堵点问题。因此标注瓶颈的解决，数据标注降价提质，是 AI产业亟需解决的问题。

（三）核心论点与研究框架

面临这些挑战，一个直观的想法就是使用人工智能技术本身来改进标注过程，而且这种尝试也确实带来了真正的革命。本文认为人工智能在数据标注过程中经历了从“助”到“主”的范式转变。这种转变不是单纯的时间优化，而是对标注过程、经济模型甚至人机协作模式的重新定义。早期以主动学习为代表的 AI技术是“助”，帮助人类标注者做数据标注。现在以基础模型为代表的新一代 AI技术是“主”，机器可以完成复杂的海量数据标注工作，人类转而成为高阶的监督管理角色。

本文将详细介绍由人工智能驱动的数据标注的研究进展，首先是传统人工标注的局限性和 AI辅助模式存在的问题；然后是基础模型驱动范式革命与潜在挑战分析；随后是新范式下系统的风险和人类—机器协作模式变革；最后是本文小结和展望。

一、传统数据标注的困境与 AI辅助模式的局限性

（一）重新审视“不可能三角”：成本、质量与规模

AI介入之前，数据标注都是手动标注，这就注定存在一个“不和谐的三角”：成本、质量、规模三者之间无法互相促进、齐头并进。人工标注的成本非常高，不仅是薪酬成本，还有随着项目数量级增加的管控成本、质量问题的返工成本、满足《个人信息保护法》等相关法律法规而产生的合规成本[8]。一些实证研究表明，企业项目数据标注过程中存在“标签噪声”“标签偏差”问题，这是项目标注员主观臆断、认知偏差、疲劳等问题导致的，即使是资深专家也无法避免[9]。可以视为是上文“标注税”的一种。在动辄数百万甚至上亿样本的数据库面前，纯人工标注的模式显得时间、资源“力不从心”，成为拖慢项目进度的瓶颈因素。

（二）“主动学习悖论”：AI辅助模式的局限性

针对传统标记难问题，学者们引入 AI当“辅助”，主动学习是最早也是最有效的策略之一[10]。主动学习的理念不在于取代人工，而是在 AI的帮助下从海量无标签数据中挖掘“更聪明”的样例，使得使用更少的标记可以取得相同的模型效果[11]。这是“聪明标记 而非更多标记”的解决方案。

主动学习并非没有成功的案例，在一些特定情况下，有的研究甚至可以证明能够减少三到六成的标注量。然而这种范式的最大内在矛盾是所谓的“主动学习悖论”。主动学习的查询函数（如不确定性采样）挑选模型最不确定的样本，即离决策边界最近的、不确定分布下的样本[12]。换句话说，主动学习的策略从整体上减少了被标注的样本数量，但与此同时，却系统地增加了每个被标注样本的平均认知难度。

这一悖论会产生两个后果：（1）只有更高水平的标注员知识与经验才能胜任困难样本标注，导致单位标注代价的升高，抵消了部分代价降低的收益。（2）对模糊的困难样本身无止境地处理可能导致标注员产生疲劳，且发生在对模型提升最有效的数据集上，标注质量风险反而增大[13]。已有大量文献与讨论认为，主动学习并非万能，过度处理困难样本可能会导致“选择忽视”，甚至会导致建模结果的局部极值。由此，主动学习范式证明了，AI对数据的“挑”并不能从根本上解决人工标注质量不佳的问题，其只是将问题从“数量”纬度转移到了“质量”“难度”纬度，从而为更加强大的、直接标注的人工智能 AI范式的出现开辟了空间。

二、基础模型引领的标注范式革命

（一）基础模型作为标注器的崛起

伴随着深度学习的迅猛发展，基础模型的诞生标志着当前 AI在数据标注领域已经从“辅助”变成了“主角”。基础模型尤其是大语言模型（Large Language Models）和视觉基础模型在大数据上进行预训练，具有超强的泛化能力和“涌现”能力[14]。它们不是筛查器，而是生成器。它们最强大的能力被称为“情境学习”（In-Context Learning）[15]。LLM可以通过在输入的指示中提供任务描述和少量示例（即“少样本学习”）后直接执行复杂标注任务，而无需对模型参数进行任何微调[16]。因此 LLM可被称为一种“通用标注器模型”，能够独自生产大数据，训练下游的“任

务学习器模型”。一项对 LLM 的数据标注的综述，也意味着这种技术范式的确立 [17]。

这种范式的转变从本质上打破并重塑了数据生产的经济模型和工作流。面向产业，智能标注平台通过整合自动预标注引擎、工作流和全链路质量管理，将数据标注瓶颈从劳动密集型的数据生产转变为模型治理和精化的认知级任务，显著提升产业效率。

（二）案例研究：Segment Anything Model (SAM)与“泛化 – 特化鸿沟”

Segament Anything Model(SAM)模型是视觉范式革命中最具代表性之一，其核心创造就是“可提示分割”任务，利用点击(键)，画框等简单提示，在零示例条件下，将任何图像中任何物体分割任务做高质量分割，使其成为划时代意义的图像标注器，就连其作者都在利用数据引擎的概念来驱动模型与数据共同进化，自动生成了10亿多个分割掩码组成了庞大的 SA-1B 数据集 [18]。

但是这种通用的泛化能力不是“无所不能”的，基础模型用在非常特化的任务领域上会有明显的效果下降，我们称之为“泛化 – 特化鸿沟”。SAM 对高对比度的自然场景图片泛化能力很好，但被用到医学图像等特化领域零样本表现会急剧下降 [19]，例如分割 x 光片中边界明确的骨骼，而对 MRI 中边界不清晰的脑肿瘤和超声图片中嘈杂的肿瘤病灶泛化效果变差 [20]。这是通用基础模型的局限性，他们的价值在于提供了强大的预训练基础，但要真正落地到高价值领域，仍需要领域自适应或微调（finetune）来消除鸿沟，如自动驾驶激光雷达点云 4D 标注、工业图像像素级瑕疵检测 [21]。而 MedSAM 就是将 SAM 在大量的医学影像数据上微调，从而拥有了较高的医学分割精度。所以，基础模型并没有消除对专业性及高质量数据的需求，只是将问题从“如何标注”转移到了“如何微调、适配”。

（三）合成数据生成：前景与陷阱

除了作为直接的标注工具，基础模型还催生了另一种强大的数据生产方式：合成数据生成 [22]。LLM 强大的生成能力使其能够创造出大量人工但高度仿真的训练样本，用于增强甚至替代真实世界数据，尤其适用于真实数据稀缺、标注成本高昂或涉及隐私敏感信息的场景。这种技术通过提示工程让模型生成特定类型的数据，或通过“模型蒸馏”用一个强大的“教师”模型生成数据来训练一个更小、更高效的“学生”模型 [23]。

但合成数据本身所具有的巨大潜力和巨大风险也同时存在。第一，事实 / 幻觉问题。LLM 可能产生出置信度很高、看似逻辑合理，但实际上完全虚构或不符合事实的“幻觉”。如果不对其进行错误标签过滤，将对训练集造成巨大损失。第二，偏见放大风险。由于生成数据的 LLM 是根据包含社会历史互联网数据进行训练的，因此，LLM 的内在偏见也会被遗传到合成数据甚至被放大。第三，基于模型自身产生的数据训练模型，可能造成“模型坍塌”，即多样性和泛化能力退化，最终陷入闭门造车的“小鸡窝”（上下文交叉污染）现象 [24]。因此，新的范式下的数据瓶颈已不是数据量而是数据本身的完整性、真实性、公平性如何被验证的扩展性问题。

三、新范式的挑战与人机协同的演进

（一）新范式的系统性风险

当数据标注的任务被大规模下放给基础模型时，模型自身存在的问题也可能被系统性地引入全局数据中。最严重的问题是算法歧视。算法歧视是指因训练数据中的缺陷、或因算法本身的主观性、或因数据代理中的错误关系导致的系统性错误。基础模型因受不平等的社会历史部分影响，它们将继承并在新的数据中扩大算法歧视。

现实中不乏这样的例子，比如亚马逊就曾因招聘中的算法歧视而不得不暂停使用。这些意味着歧视并非技术本身的问题，而是人类社会与人类历史的现实。当一个带有偏见的 LLM 作为标注器大量生产带有偏见的数据时，就形成了一个自我强化的“反馈循环”，会让下游的应用程序在招聘、司法甚至金融等影响社会生活的关键领域做出不平等的决定，加剧社会不平等 [25]。于是，如何在大规模自动化生产的过程中检测、测量、减少偏见就成了这种新范式下的伦理与技术问题。

（二）人机协同的演变：从标注员到监督者

当然，基础模型只是作为主导力量进行标示，并不意味着人的角色的消失，只是人机协作的形态发生了根本性改变，人从简单重复性劳动的“劳动者”变成了对高阶认知任务的“监督”和“管理”，这种新的协同就是新的“人机共生”。

在这个新的范式中，人类的主要任务从标注标签转向对 AI 系统的方向指导、质量把控与价值校正。它们包括：

●提示工程与数据策展：专家需要精心设计提示，引导基础模型正确理解任务，产生高质量标注和提示生成，同时，专家还需要策展少量的高质量种子数据作为“黄金标准”用于训练和评测模型。

●验证和异常处理：人类不再需要检查每一个标签，只需专注于模型置信度最差的情况和存在高风险决策和模型无法处理的情况。

●伦理监督与价值对齐：这是新一代人机协同最根本的责任。人类通过“基于人类反馈的强化学习”(Reinforcement Learning from Human Feedback)等机制，提供反馈校准基础模型，对齐人类价值、安全、伦理 [26]。未来产业能否具有强大的竞争力，取决于人类能否在技术、道德、领域等方面对模型有更适配的限制和调教。

（三）可持续性：一个未来的研究方向

随着基础模型越来越大，训练运行这些基础模型需要消耗巨大的算力资源、计算资源。这会产生巨大的经济成本与环境影响。探求“绿 AI”、绿色可持续数据标注范式，是未来的研究重点。

未来的发展可以探索在不必牺牲性能的前提下，降低 AI 系统的碳足迹的方法，使用节能技术、可再生能源数据中心，通过转向“以数据为中心的人工智能”[27]，后者提倡用更小的、高质量、精致的训练数据集训练模型来替代大量消耗的计算资源，也更符合我国走高质量发展之路的思路，这为数据标注的“成本”维度

增添新的内容，其能否在未来获得成功，取决于能否构建一个高效的可持续的 AI 生态系统。

四、结论与展望

数据标注由人工作业化，发展到 AI 辅助作业，再到基础模型带动的全新阶段。这个变化的实质，是 AI 由协助数据生产的“小工”，到直接“涌现”数据的“大工”的角色转变。这不仅仅是技术本身的发展，还有数据生产的人机关系的变化。

在新的范式下，AI 完成标注任务，人从一线劳动者的身份，变成 AI 系统的指挥官、监督员、价值校准器，这种新形态的“人机共生”是新一代 AI 的显著特征。

面向未来，数据标注领域的发展方向不是“无人”，而是一个更完备、更高效的生态系统，在这样一个生态系统中：基础模型是自动化的，人类的智慧、特长、知识、美德被精准地使用到定义规则、处理模糊边界案例、纠正系统偏见中，使未来的人工智能永续造福人类。这种深度、高密、互生共长的人机关系是未来人工智能发展的重要方向，也是未来人工智能健康、可持续发展的保障。

参考文献

[1] Zhu X, et al. 数据标注研究综述 [J]. 软件学报 , 2020, 31(1): 1-20.

[2] Neves M, Ševa J, Teixeira L. On the challenges of using text annotation tools[C]//Proceedings of the 12th Language Resources and Evaluation Conference. 2020: 5353-5360.

[3] 中国信息通信研究院 . 人工智能数据标注白皮书 [R]. 2023.

[4] 国务院办公厅 . 关于促进数据标注产业高质量发展的实施意见 [Z]. 2024.

[5] 王珊 , 李建中 . 数据库管理系统中的不确定性数据管理研究进展 [J]. 软件学报 , 2018, 29(1): 1-20.

[6] Midtvedt B. Annotation-free deep learning for quantitative microscopy[D]. University of Gothenburg, 2024.

[7] Sun S, Liu L, Liu Y, et al. Uncovering bias in foundation models: Impact, testing, harm, and mitigation[J]. arXiv preprint arXiv:2501.10453, 2025.

[8] 工业和信息化部 . 人工智能产业创新发展行动计划 (2023-2025 年) [Z]. 2023.

[9] Settles B. Active learning literature survey[R]. University of Wisconsin-Madison, 2009.

[10] 周志华 . 机器学习 [M]. 北京 : 清华大学出版社 , 2016.

[11] 李飞飞 , 张钹 , 等 . 人工智能的本质与未来 [J]. 中国科学 : 信息科学 , 2024, 54(1): 1-25.

[12] Dasgupta S. Analysis of a simple active learning algorithm[C]//Proceedings of the twenty-first conference on Uncertainty in artificial intelligence. 2005: 133-140.

[13] Beygelzimer A, Dasgupta S, Langford J. Importance weighted active learning[C]//Proceedings of the 26th annual international conference on machine learning. 2009: 49-56.

[14] Brown T B, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901.

[15] Min S, Lyu X, Holtzman A, et al. Rethinking the role of demonstrations: What makes in-context learning work?[J]. arXiv preprint arXiv:2202.12837, 2022.

[16] Zhao W X, Zhou K, Li J, et al. A survey of large language models[J]. arXiv preprint arXiv:2303.18223, 2023.

[17] Tan Z, Beigi A, Wang S, et al. Large language models for data annotation: A survey[J]. arXiv preprint arXiv:2402.13446, 2024.

[18] Kirillov A, Mintun E, Ravi N, et al. Segment anything[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 3879-3890.

[19] Mazurowski M A, Dong H, Gu H, et al. Segment anything model for medical image analysis: An experimental study[J]. Medical Image Analysis, 2023, 89: 102918.

[20] Ma J, He Y, Li F, et al. Segment anything in medical images[J]. Nature Communications, 2024, 15(1): 654.

[21] Zhang Y, Liu Q, Song L. Medical image segmentation based on U-Net and its variants: A review[J]. Journal of Shanghai Jiaotong University (Science), 2023, 28(3): 345-358.

[22] Veselý A, Straka M. Synthetic data generation using large language models: A survey[J]. arXiv preprint arXiv:2503.14023, 2025.

[23] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[J]. arXiv preprint arXiv:1503.02531, 2015.

[24] Shumailov I, Zhao Y, Bates D, et al. The curse of recursion: Training on generated data makes models forget[J]. arXiv preprint arXiv:2305.17493, 2023.

[25] 段伟文 . 人工智能伦理与治理 [M]. 北京 : 科学出版社 , 2023.

[26] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback[J]. Advances in Neural Information Processing Systems, 2022, 35: 27730-27744.

[27] 张俊林 . 以数据为中心的人工智能：现状与展望 [J]. 中国科学 : 信息科学 , 2023, 53(8): 1459-1476.

