机器学习画布：为何战略比人工智能代码生成更重要的实证发现

Martin Prause

摘要 尽管人工智能编程助手日益普及、超过 80% 的机器学习（ML）项目仍无法创造真正的商业价值。本研究创建并检验了

索引词— 大型语言模型，机器学习，项目管理，软件工程，软件开发管理

I. 引言

实证研究表明，大型语言模型（LLM）能够通过自动化某些编码任务来提升开发人员生产力。例如，使用 Copilot 的开发人员完成任务的速度最多可提升 55%，认知负担也更轻 [1]。这对经验较少的程序员尤其有益。Anthropic 对约 50 万次编码会话的分析显示，与 Claude Code 的交互中，约有 79% 涉及自动化 [2]。一项实证研究表明，人工智能提高了 Python 编程任务的生产力；不过，研究人员提醒，其效果取决于任务复杂度和用户的专业知识 [3]。此外，国际清算银行开展的一项大规模现场实验显示，生产力提高了 55%，大语言模型生成的代码约占代码总行数的三分之一 [4]。尽管这些发现证实了生产力的提升，但也表明人工监督仍不可或缺。人工智能（AI）生成的代码主要用于处理重复性或定义明确的编程活动，而非复杂的架构决策 [5]。

尽管大语言模型能够加快代码生成，并减轻日常编程任务带来的认知负担，但这种微观层面的效率提升并未转化为宏观层面的项目成功。一项针对 65 位资深数据科学家的研究发现，超过 80% 的人工智能项目以失败告终，是传统信息技术项目失败率的两倍 [6]。作者指出了五种主要失效模式：（1）技术目标与业务问题不匹配；（2）数据质量和基础设施不佳；（3）试图解决超出当前人工智能能力范围的问题；（4）组织准备度不足；（5）治理结构不完善。这些发现证实了早期研究的结论，即机器学习（ML）部署的失败率很高 [7][8]。

为应对这些挑战，机器学习运维（MLOps）应运而生，成为一门将软件工程原则与机器学习特定要求相结合的学科 [9]。一项系统性文献综述指出，机器学习运维涵盖持续集成与部署、自动化测试与验证、模型版本控制与治理，以及生产监控与维护 [10]。同样，另一项研究表明，采用机器学习运维的成功率因MLOps成熟度不同而有显著差异。成熟度较高的企业在数据管理、自动化部署以及持续集成与交付方面展现出显著更强的能力 [11]。

这些多方面的挑战凸显了采用结构化框架来规划、沟通和执行人工智能与机器学习项目的必要性。由Osterwalder推广的商业模式画布（BMC）方法已被证明

Martin Prause, martin.prause@xinblue.de。

有助于在不同的利益相关者 [12]之间建立共同的心智模型。近期在人工智能领域的应用包括人工智能模型画布，它将画布原则调整为满足机器学习的特定要求[13]。然而，现有的画布框架往往未能涵盖机器学习项目的全部复杂性。它们通常只关注技术方面（如数据管道和模型架构）或业务考量（如价值主张和成本结构），却未能有效整合这两种视角。此外，这些框架对机器学习项目的动态特性关注不足，而这类项目需要灵活且结构化的方法来开展迭代实验[14]。

II. 背景：机器学习画布

为应对这些局限，本文开发了机器学习画布（MLC），将组织理论、软件工程和数据科学融为一体。研究采用结构方程模型（SEM）方法进行实证分析，以识别在大型语言模型充当编程助手的开发情境中，机器学习项目的成功决定因素。

商业模式是一份蓝图，使公司的战略目标与运营执行保持一致 [15]。它阐释了组织如何创造、交付并获取价值 [16]。层级分类体系确定了商业模式的四个基本维度：1. 价值主张：创造什么价值，以及为谁创造？2. 价值架构：如何通过组织能力创造价值？3. 价值网络：促成价值创造的关系生态系统。4. 价值财务：获取所创造价值的经济模式。

商业模式画布通过将九个相互关联的构成模块排列在可视化画布上，对这些维度进行运营化。这种方法有助于迭代式开发，使利益相关者能够在处理具体组成部分的同时保持整体视角。机器学习项目与商业模式具有结构上的相似性，因为二者都通过协调多个视角，专注于从数据中创造价值：商业战略、技术实施和运营整合。这些特征要求建立一个专门的框架，以处理机器学习计划的技术和组织层面。

表一 机器学习画布

问题	模型	维度	画布元素
创造了什么价值?	BMC	价值主张	价值主张
	MLC	战略	任务、数据、目标
如何创造价值?	BMC	价值架构	关键活动
	MLC	流程	数据收集, Data 准备、算法选择、训练与评估、性能优化、风险
需要什么才能创建价值?	BMC	价值基础设施	关键资源、关键合作伙伴，客户关系，客户细分、渠道
	MLC	生态系统	部署基础设施，开发基础设施，生产集成
驱动因素业务影响创造的价值?	BMC	价值财务	成本结构，收入结构
	MLC	支持	项目管理与治理 组织领导力

III. 方法论

为了考察机器学习项目成功的决定因素，并建立一个基于理论的模型，我们将 BMC 维度映射到相应的机器学习构念上。这些构念包括战略、流程、生态系统和支持。接下来，我们从文献中为每个维度确定了自变量测量指标。

因变量是 IT 项目成功，通过铁三角框架进行测量，该框架由绩效目标、业务目标和利益相关者满意度构成[17]。尽管数据科学家经常关注技术绩效指标，例如 AUC 分数或模型准确率，但铁三角采用的是业务利益相关者能够理解的语言：时间、成本和质量。尽管机器学习项目具有技术复杂性，但从根本上说，它们是必须在约束条件内交付价值的业务投资。如果一个模型达到了 99% 的准确率，但交付延迟了六个月，且成本超出预算一倍，那么这一准确率就变得毫无意义。铁三角通过衡量利益相关者所关心的事项来体现这一现实：项目是否按时交付了其承诺的功能，并且在预算

get.

表二 IT 项目成功的构念测量指标

Item	您参与的机器学习项目通常被认为是成功的，如果.....
S1_PERF	机器学习模型达到绩效目标。
S2_BO	机器学习项目的结果符合业务目标。
S3_STAK	所有利益相关者都对机器学习项目感到满意。

自变量由机器学习画布的不同维度构成，这些维度采用5点李克特量表进行测量，量表范围从“非常同意”到“非常不同意”。结构方程模型方法能够检验机器学习画布的四个维度——战略、流程、生态系统和支持——如何共同作用并影响项目成功。该方法论承认，这些维度无法直接观测，而是通过多个调查题项测量的潜在构念。

A. 战略维度

机器学习战略维度将业务目标转化为机器学习问题[18]。首先，任务定义要素要求将业务问题映射到适当的学习范式，以识别适合机器学习的任务

[19]。这包括：1）区分预测、探索和优化问题；2）评估数据可用性；以及 3）为结果定义可解释性要求 [20]。

在启动任何机器学习项目之前，数据规范是一项必要工作。它规定数据源和模态，并区分结构化与非结构化数据，以及如何存储和处理这些数据 [21]。质量考量如今已不再局限于传统的数据库完整性，还包括代表性、时间稳定性以及可能损害模型有效性的潜在偏差 [22]。近期研究强调，数据级联（即数据质量不佳导致的错误不断累积）是机器学习项目失败的主要原因 [23]。

目标定义将技术性能指标与商业价值创造联系起来。这有助于解决机器学习系统中隐性的技术债务 [7]。定量指标不仅应涵盖准确率或 F1 分数，还应包括计算效率、延迟约束以及对分布偏移的稳健性 [24]。由《通用数据保护条例》（GDPR）等监管框架所推动的可解释性要求，以及防止歧视性结果的约束条件等定性考量，也日益同等重要 [25][26]。负责任人工智能框架的兴起，要求在整个机器学习生命周期中明确考虑透明度、问责制和可审计性 [27]。

这些概念要素将战略维度转化为可测量的条目，用于我们的实证分析
ysis.

表三 战略维度的构念测量指标

Item	战略维度
ST1_目标	机器学习项目的成功衡量指标已明确定义。
ST2 任务_	机器学习项目中的任务定义清晰。
ST3 数据_	机器学习项目的要求，例如数据和模型约束条件（可解释性、空间、推理时间等）定义清晰。

B. 流程维度

机器学习流程维度采用经过调整的 CRISP-DM 框架，将原始数据转化为可部署的模型。该框架承认机器学习开发具有迭代性和非确定性 [28]。

数据收集需要采用提取-转换-加载（ETL）流程，以应对大数据的“5V”：规模、多样性、速度、真实性和价值 [29]。规模涵盖存储与治理架构，这些架构需要在可访问性与安全性之间取得平衡；也涵盖分布式系统，它们确保可扩展性，同时不牺牲一致性 [30]。数据湖必须同时容纳结构化与非结构化数据，并保持可追溯性和版本控制。多样性要求通过语义映射和模式匹配等方式，构建能够协调异质数据的集成流水线，例如关系模式、非结构化文本和时间流 [31]。为此，需要稳健的编码策略：类别变量需要采用独热编码、目标编码或嵌入方法，这些方法会影响可解释性和维度；文本需要进行向量化和嵌入；时间特征则需要采用周期编码，以表示季节性模式。速度带来了时间复杂性：批处理与流处理并存，在特征计算和模型更新方面各有不同的约束条件 [32]。标准化和归一化会影响模型收敛，并在异常值敏感性和信息保留方面带来不同的权衡。质量评估还涵盖机器学习特有的考量，例如标签噪声容忍度、特征漂移检测和表示充分性 [33]。多维度质量——包括完整性、一致性、及时性和相关性——需要通过综合指标来衡量，并体现下游性能，而非孤立的特征 [34]。

数据准备最多会耗费项目 80% 的工作量 [35]。清洗流程必须在信息保留与噪声降低之间取得平衡。例如，移除异常值可能会删去罕见但合理的模式，而这些模式对于预测罕见事件至关重要 [36]。现代方法采用弱监督和程序化标注，拓展人类专业知识的应用范围。这些方法将领域知识编码为标注函数，而非人工标注 [35]。特征工程是连接领域知识与统计学习的重要接口。实践者必须在自动化机器学习（AutoML）或深度学习架构的自动特征学习能力，与运用统计方法或通过可视化探索特征重要性来实现的手工设计特征的可解释性和效率之间取得平衡 [37]。特征选择与转换流水线体现了不断提高的抽象层次。选择方法（过滤式、包装式和嵌入式）优化现有特征子集。提取技术（例如降维）用于发现潜在表示。转换（例如归一化、离散化和编码）则确保算法兼容性 [38]。

算法选择不只是找到一个达到性能目标的优秀算法。无免费午餐定理将这一点形式化：没有任何单一算法能在所有问题领域中占据优势 [39]。这一基本约束要求采用组合方法，即算法的选择应综合考虑问题特征（例如数据分布、噪声水平和特征类型）、运行约束（例如延迟、内存和可解释性）以及组织能力（例如专业知识和维护能力）。

神经架构搜索和自动化机器学习的兴起，使挑战从在固定架构之间进行选择转变为探索大型配置空间，尽管这有时会以牺牲可解释性为代价 [40]。迁移学习将范式从从头学习转变为适应已有知识。在这一方法中，来自大规模数据集的预训练表征充当

初始化点，从而降低 后续任务[41] 的样本复杂度。集成方法通过装袋法降低方差、提升法降低偏差，或堆叠法进行层级组合，借助多样性实现稳健性，从而将“群体智慧”原则付诸实践 [42]。

训练和评估流程必须在多个相互竞争的目标之间取得平衡，包括统计有效性、计算效率和业务相关性。训练数据的划分策略包括随机划分，以及针对时间序列数据的时间划分、针对类别不平衡的分层抽样，以及针对层次数据结构的分组划分。每种策略都能保留数据生成过程的不同方面。设计目标函数是将业务目标转化为数学优化的重要步骤。替代损失在保持可微性和计算可处理性的同时，近似真实的业务指标 [43]。现代评估框架强调多个维度上的稳健性，例如不同数据子群间的性能稳定性（切片评估）、分布偏移下的性能退化（压力测试），以及不同随机种子下的一致性（方差分析）。

性能优化受模型容量、训练数据和计算资源构成的约束三角关系制约。在某一方面取得改进，往往需要在其他方面作出牺牲。错误分析已从单纯检查混淆矩阵，发展为运用错误分析循环和反事实调试等复杂框架。这些框架会系统地探查模型失效，以识别可修正的模式 [44]。偏差-方差分解理论为优化策略奠定了基础，例如通过提高模型容量或进行特征工程来解决高偏差问题，或通过使用正则化、数据增强或集成方法来应对高方差问题 [45]。超参数优化已从网格搜索发展到能够对响应面建模的贝叶斯优化方法。近期的多保真优化进展大幅降低了计算成本 [46]。

机器学习项目中的风险评估聚焦于机器学习方法所特有的失效模式 [47]。这些失效模式包括数据偏差和算法偏差，以及模型脆弱性与安全性。数据偏差可能表现为训练数据中编码的历史偏见、数据选择中的偏差，以及特征定义中的偏差。每种偏差都需要不同的缓解策略，范围从重新加权到因果去偏 [48]。算法偏差源于学习算法固有的归纳偏置，例如正则化模型偏好简单假设、最近邻方法的局部性偏置，以及深度网络的层次偏置。这些偏差可能以复杂的方式放大或抵消数据偏差 [49][50]。对抗脆弱性揭示了机器学习系统易受恶意输入影响；难以察觉的扰动可能导致灾难性的错误分类。因此，需要采取防御策略，范围从对抗训练到认证鲁棒性 [51]。这些风险相互交织，形成复合脆弱性，例如，存在偏差的训练数据可能助长定向对抗攻击，算法偏好则可能造成可预测的失效模式。这要求采用整体风险框架，考量各种风险之间的相互作用，而非将威胁孤立看待 [52]。

流程维度通过关键的机器学习活动进行操作化。我们区分预处理活动（如数据收集、清洗和特征工程）与后处理活动（如算法选择、训练、评估、优化和风险评估）。我们认识到，这些阶段需要不同的技能组合和组织能力。

表 IV 流程维度的构念测量指标

Item	**流程维度：预处理**
PRE1_COLL	收集和整合数据是一项简单直接的任务。
PRE2 清洗_	清洗和准备数据是一项简单的任务。
PRE3 FEAT_	从数据中识别合适的特征是一项简单的任务。
Item	**流程维度：后处理**
POST1 风险_	评估机器学习模型的潜在风险是一项简单直接的流程。
POST2 训练	训练机器学习算法是一项简单直接的流程。
帖子3 OPTI	优化机器学习模型是一个简单明了的流程。
帖子4 选择_	选择并测试不同的机器学习算法是一项简单明了的任务。
帖子5 评估_	评估机器学习模型是一项简单直接的任务。

C. 生态系统维度

机器学习生态系统维度指的是将机器学习实验转化为生产系统的社会技术基础设施。尽管机器学习算法仅占现实世界中的机器学习系统的 5%，但支持性基础设施决定了模型能否创造价值，还是沦为成本高昂的实验 [53]。这一生态系统必须协调相互竞争的需求：机器学习开发具有探索性，需要灵活性和快速迭代；而生产要求则强调可靠性、可复现性和治理。

机器学习项目的基础设施架构已从临时拼凑发展为复杂的平台。基础层整合了各种数据源，包括业务数据库、数据仓库、流处理平台和应用程序编程接口 [54]。计算层的专用硬件，从图形处理器到张量优化型张量处理器，为工作负载调度带来了挑战。安全层应对机器学习特有的问题，例如保护训练数据的差分隐私、防范模型提取，以及监管审计追踪记录 [55]。尽管云平台提供弹性扩展、预配置服务以及使用尖端硬件的机会，但数据重力、监管约束和延迟要求使得混合架构乃至本地部署资源成为必要 [56][57]。成本优化

需要采用细致的策略，例如在训练时使用竞价实例，并为推理预留容量。多云策略有助于避免供应商锁定，并在风险缓解与运营复杂性之间取得平衡 [58]。

机器学习开发工具的生态系统需要整合统计计算环境、软件工程工具链和领域专用平台。编程语言的选择体现了根本性的权衡。例如，Python 生态系统的主导地位和易用性，与编译型语言的性能优势之间就存在取舍。这通常会催生多语言架构，其中 Python 充当编排层 [59]。机器学习框架本身也体现了不同的理念：TensorFlow 以生产为先，采用静态计算图；PyTorch 面向研究，支持动态计算；JAX 则采用函数式编程范式，从而实现新颖的优化策略。实验跟踪和版本管理工具有助于应对机器学习领域的可复现性危机：它们记录的不仅是代码，还包括超参数、随机种子和计算环境 [60]。协作平台弥合了数据科学家偏好的基于笔记本的探索与工程师偏好的基于集成开发环境的开发之间的差距，同时支持同行评审 [61]。监控和调试工具面临机器学习领域独有的挑战。例如，检测数据漂移、模型退化和公平性违规，需要具备统计学方面的专业能力，远超传统应用性能监控所需的水平 [62]。

生产集成是机器学习项目中的“几乎最后一公里”难题：由于工程方面的挑战，前景良好的原型往往无法实现其价值。技术集成需要审慎进行 API设计，以应对机器学习预测固有的不确定性；需要采用版本管理策略，以支持逐步发布和回滚；还需要采取容器化方法 [63]。在机器学习生命周期的不同阶段，可扩展性挑战各不相同。训练可扩展性通过数据并行和模型并行实现。服务可扩展性通过批处理和缓存策略实现。特征计算可扩展性则通过增量处理实现 [64]。模型监控包括跟踪关键指标，并通过统计检验检测行为漂移、通过不确定性量化检测对抗性输入，以及对不同人口群体开展公平性监控 [65]。通过低代码平台和自动机器学习服务推动机器学习民主化，有望提升其可及性，但同时也引发了对抽象与控制之间权衡的担忧，以及对可能形成“黑箱”基础设施的担忧——这种基础设施会重现模型本身的复杂性 [66]。

表 IV 生态系统维度的构念测量指标

Item	生态系统维度
ECO1 DEV_	机器学习项目期间的开发与协作是一项简单直接的任务。
ECO2 INT_	在生产环境中集成机器学习模型是一项简单的任务。
ECO3 基础设施_	使用所提供的 IT 基础设施是一项简单的任务。

D. 支持维度

支持维度关注人工智能采用中“组织学习”的关键要求：认识到，若缺乏相应的组织能力和治理结构，仅凭技术卓越并不能确保机器学习项目成功 [67]。该维度涵盖管理和领导实践，旨在弥合机器学习潜力与实际商业价值之间的“部署鸿沟”。这一鸿沟导致多达 87% 的数据科学项目止步于生产部署之前 [68]。

管理机器学习计划需要在数据科学的探索性与组织对可预测性和控制的要求之间寻求平衡。机器学习项目需要由混合团队组成，融合领域专业知识、统计知识、工程能力和商业敏锐度 [69]。为适应机器学习固有的不可预测性，开发方法必须融入敏捷原则，由此形成“机器学习冲刺”等方法，在更广泛的交付框架中嵌入实验周期 [70]。

要实现机器学习项目成功，组织领导层必须推动企业能力多个维度的转型。利益相关者参与必须应对“人工智能信任赤字”，即公众对人工智能公平性和透明度的怀疑，与员工对自动化和岗位替代的恐惧之间的交集。这需要采取沟通策略，强调增强而非取代、透明而非保密 [71]。绩效衡量还必须体现商业价值。这需要制定“人工智能转型指南”，通过运营改进、客户满意度和战略定位来界定成功，而不只是看模型准确率和量化性能 [72]。组织必须培养“人工智能素养”，即识别机器学习机会、管理机器学习项目并将机器学习输出整合到决策流程中的组织能力 [73]。

项目管理与组织领导力的交汇，体现在兼顾创新与责任的治理结构中。机器学习治理框架必须应对传统 IT 治理中不存在的独特挑战，包括模型生命周期管理、版本管理与弃用政策、针对受保护属性开展公平性与偏差监控、针对高风险决策制定可解释性要求，以及持续监控性能退化 [74]。欧盟《人工智能法案》等监管框架引入了外部治理要求，包括风险评估、透明度措施，以及对高风险人工智能应用实施人工监督。这使合规从一份技术核对清单转变为一项战略要务 [75]。

这些业务支持要素可归纳为三项

组织承诺的基本指标：机器学习项目是否获得公司范围内的认可（SP1_WELL）、是否被列为战略优先级（SP2_PRIO），以及是否获得充足的财务资源（SP3_FIN）。这些可观察的标志反映了组织将所述复杂支持要求内化并付诸行动的程度。

表 V 支持维度的构念测量指标

Item	生态系统维度
SP1_WELL	机器学习项目在公司内受到广泛认可
SP2 PRIO_	机器学习项目在公司内具有很高的优先级。
SP3 FIN_	机器学习项目的财务预算充足。

IV. 建模

以下模型的结构体现了一个基本原则：机器学习项目具有独特特征，因此需要特定的组织能力和顺序依赖关系 [7][14]。该模型的结构回应了机器学习项目中探索与利用之间的根本张力 [76]。组织支持通过资源供给推动探索，战略引导探索朝着特定目标发展，流程通过技术实施利用选定的方法，而生态系统则将成功模式制度化。这种从探索到利用、逐步提高具体程度的演进，解释了为何其他因果结构会违背机器学习开发的内在逻辑。

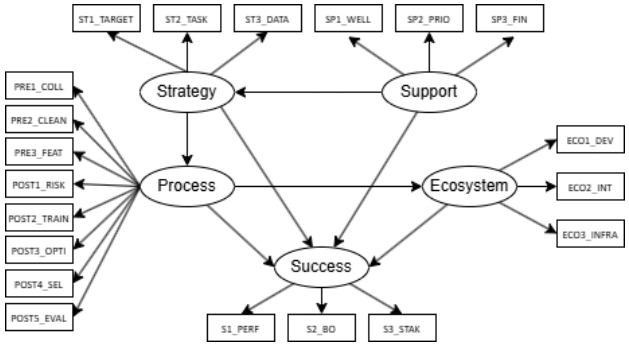

图1. 基于机器学习画布各维度的结构方程模型，反映各构念之间的因果依赖关系。

将组织支持定位为初始因果驱动因素，与资源基础观理论相符；该理论认为，组织资源和能力决定战略选项 [77]。这种因果关系不仅仅是相关关系；实验性证据表明，组织支持的变化会直接影响数据驱动型战略举措的质量和范围 [78]。支持与下游要素之间不存在直接路径，这反映了“分析鸿沟”——没有战略转化，仅凭资源无法创造价值 [79]。

战略作为支持与流程实施之间中介的核心作用，在理论上有着坚实的基础，源于战略协同文献。IT战略是连接组织能力与技术执行的关键桥梁 [80]。对于机器学习项目而言，战略清晰度——通过任务定义、数据规范和目标设定加以具体化——决定了技术开发的整体轨迹 [81]。

从战略到流程的因果链反映了机器学习开发的内在逻辑，即战略决策先于技术实施。采用战略优先方法的机器学习项目，其成功率显著高于从技术试验开始的项目[82]。

在机器学习环境中，部署基础设施应源于技术流程决策，而不是决定这些决策[8]。谷歌的机器学习工程实践为这一因果方向提供了实证支持：团队会先建立数据管道和模型架构，然后再设计部署系统[14]。

V. 结果

为检验这种结构化方法是否能促成成功的机器学习项目，我们对数据科学家进行了调查。该调查在向3,000名数据科学家发放前进行了两轮试点测试，最终收到161份回复（回复率为5%）。清洗数据后，我们认定其中150份回复适合进行分析；这些数据科学家每天都会使用人工智能编程助手（平均每天1至2小时）。大多数受访者为在职专业人士（67%），其次是学生（20%）、自由职业者（11%）和企业主（9%）。其中40%在拥有500名以上员工的大型企业工作，其余受访者分别来自员工人数少于20人的公司（15%）、拥有20至50名员工的公司（6%）、拥有50至100名员工的公司（8%）以及拥有100至500名员工的公司（13%）。另有19%选择了“不适用”。

参与者来自多个领域，包括金融科技、教育、人工智能（AI）、在线零售、金融服务、分析咨询和电信。在这些参与者中，只有14%表示其机器学习项目产生了实际业务影响。22%倾向于认为项目具有业务影响，25%持中立态度，39%将其工作描述为实验性项目（23%倾向于实验性，16%则纯粹是试点项目）。这一分布体现了人们对机器学习采用谱系的不同看法，范围从实验性举措到生产部署。

在进行结构方程模型分析之前，我们验证了所有基本假设均已满足。首先，采用马氏距离结合卡方分布检验评估多元正态性。由于所有 p 值均超过 0.001 阈值，未发现离群值，证实了多元正态性假设。多重共线性诊断未发现问题：所有方差膨胀因子值均低于 10，容差值均高于 0.1，表明预测变量之间不存在共线性。

表VI 容差值和方差膨胀因子

模型	非标准化系数B	Std. 误差	标准系数 Beta	t	Sig.	容差	VIF
（常量）	90.72	22.792		3.98	0		
S1_PERF	-6.477	7.888	-0.094	-0.821	0.413	0.531	1.882
S2_BO	4.612	7.156	0.075	0.645	0.52	0.521	1.921
S3_STAK	4.811	6.458	0.08	0.745	0.458	0.607	1.646
ST1_TARGET	-0.46	4.789	-0.011	-0.096	0.924	0.553	1.809
ST2_TASK	-2.115	5.499	-0.051	-0.385	0.701	0.403	2.482
ST3_DATA	-6.998	5.373	-0.161	-1.303	0.195	0.454	2.205
SP1_WELL	6.084	4.922	0.142	1.236	0.219	0.527	1.899
SP2_PRIO	-5.894	5.005	-0.136	-1.178	0.241	0.526	1.902
SP3_FIN	2.997	4.917	0.065	0.61	0.543	0.619	1.615
PRE1_COLL	-9.476	6.303	-0.223	-1.503	0.135	0.315	3.171
PRE2_CLEAN	15.531	6.474	0.362	2.399	0.018	0.306	3.268
PRE3_FEAT	-3.156	5.694	-0.069	-0.554	0.58	0.455	2.195
POST1_RISK	4.513	6.38	0.096	0.707	0.481	0.38	2.629
POST2_TRAIN	-4.781	5.7	-0.102	-0.839	0.403	0.467	2.143
POST3_OPTI	1.383	5.876	0.031	0.235	0.814	0.396	2.527
POST4_SEL	-1.418	5.641	-0.028	-0.251	0.802	0.553	1.807
POST5_EVAL	-0.576	7.132	-0.012	-0.081	0.936	0.309	3.241
ECO1_DEV	-3.199	6.169	-0.069	-0.519	0.605	0.388	2.575
ECO2_INT	-3.197	5.218	-0.077	-0.613	0.541	0.439	2.28
ECO3_INFRA	3.348	6.148	0.077	0.545	0.587	0.349	2.864

容差值（均高于 0.30）和方差膨胀因子（VIF）得分（均低于 5.0）表明多重共线性处于可接受水平，证实 SEM 关于自变量线性无关性的假设已满足。

通过采用 LOESS平滑的残差图检验了异方差性，结果呈现直线型模式，证实了同方差性。相关矩阵显示出正定性，且行列式大于零，满足结构方程模型估计的一项关键假设。

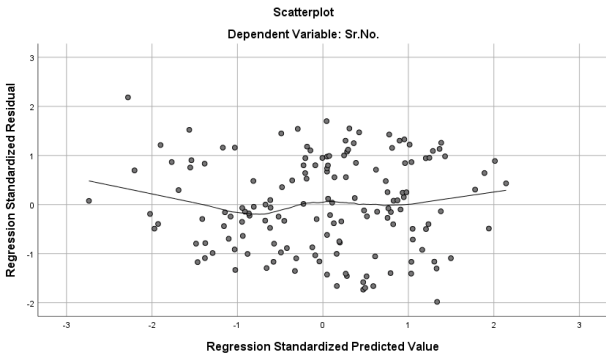

图2. 异方差性/同方差性分析

初始因子载荷显示，财务支持条目（SP3_FIN）存在过高的交叉载荷，损害了区分效度。按照既定程序，我们在后续分析中剔除了该条目，从而改进了模型设定。信度分析表明，所有构念均具有较强的内部一致性，各维度的克朗巴赫α系数均超过 0.7 的阈值。修正后的模型如下：

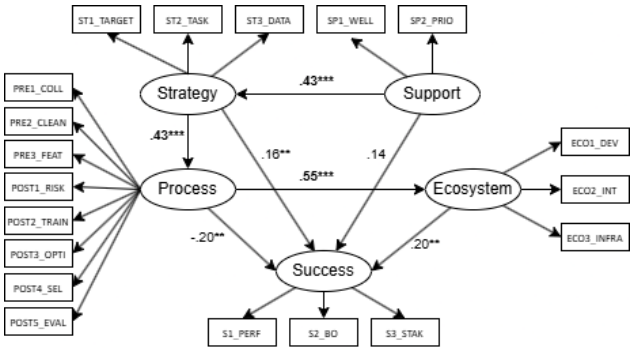

图3. 机器学习项目成功决定因素的精炼模型，不含 SP3_FIN 测量。***表示在1%水平上显著，**表示在5%显著性水平上显著。

表VII 精炼模型的回归权重

因变量	自变量	估计值	S.E.	C.R.	P值标签
战略	支持	0.432	0.12	3.601	***
流程	战略	0.428	0.086	5.006	***
生态系统	流程	0.547	0.092	5.93	***
成功	战略	0.163	0.081	2.021	0.043
成功	支持	0.137	0.076	1.809	0.071
成功	生态系统	0.202	0.08	2.524	0.012
成功	流程	-0.199	0.09	-2.217	0.027
ST1_TARGET	战略	0.783	0.101	7.769	***
ST2_TASK	战略	1			
ST3_DATEN	战略	0.898	0.102	8.778	***
SP1_WELL	支持	1			
SP2_PRIO	支持	0.829	0.185	4.471	***
ECO3_INFRA	生态系统	1			
ECO2_INT	生态系统	0.876	0.098	8.972	***
ECO1_DEV	生态系统	0.697	0.089	7.823	***
POST4_SEL	流程	0.899	0.093	9.68	***
POST3_OPTI	流程	0.838	0.088	9.49	***
POST2_TRAIN	流程	1			
POST1_RISK	流程	0.982	0.087	11.296	***
PRE3_FEAT	流程	0.863	0.087	9.909	***
PRE2_CLEAN	流程	0.939	0.093	10.088	***
PRE1_COLL	流程	0.909	0.096	9.47	***

S3_STAK	成功	0.834	0.13	6.399	***
S2_BO	成功	1			
POST5_EVAL	流程	0.919	0.098	9.341	***
S1_PERF	成功	0.902	0.13	6.958	***

验证性因子分析（CFA）证实，所有构念均具有较强的聚合效度。平均方差提取量（AVE）介于 0.58 至 0.74 之间，明显高于建议阈值 0.50，证实了较强的聚合效度。组合信度（CR）值始终较高（0.84 至 0.92），超过普遍认可的 0.70 基准，表明内部一致性极佳。因子载荷（CR 的平方根）介于 0.92 至 0.96 之间，表明各构念与其观测指标之间存在稳健关联。此外，误差方差按（1 - CR）计算，保持在较低水平（0.08 至 0.16），进一步凸显了测量模型的高可靠性与精确度。

表 VIII CFA 的结果

测量	成功	流程	生态系统	支持	战略
AVE	0.63	0.58	0.64	0.74	0.63
CR	0.84	0.92	0.84	0.85	0.84
因子载荷	0.92	0.96	0.92	0.92	0.92
误差方差	0.16	0.08	0.16	0.15	0.16

采用 Fornell-Larcker 准则确认了区分效度。每个构念的平均方差提取量（AVE）均高于其与其他构念的相关系数平方。最高的相关系数平方（0.26）出现在生态系统与流程之间，且仍低于这两个构念各自的平均方差提取量，证实每个构念均能捕捉到不与其他构念共享的独特方差。

区分效度分析（DVA）的结果显示，最高相关系数平方（0.26）出现在生态系统与流程之间，且低于这两个构念各自的平均方差提取量（AVE），证实每个构念都能捕获不与其他构念共享的独特方差。

DVA 结果表九

			因子相关性	相关性平方	AVE1	AVE2
战略	<-->	流程	0.467	0.22	0.63	0.58
支持	<-->	流程	0.184	0.03	0.74	0.58
生态系统	<-->	流程	0.507	0.26	0.64	0.58
流程	<-->	成功	0.037	0.00	0.58	0.63
战略	<-->	支持	0.427	0.18	0.63	0.74
战略	<-->	生态系统	0.315	0.10	0.63	0.64
战略	<-->	成功	0.31	0.10	0.63	0.63
支持	<-->	生态系统	0.145	0.02	0.74	0.64
支持	<-->	成功	0.314	0.10	0.74	0.63
生态系统	<-->	成功	0.26	0.07	0.64	0.63

模型拟合度评估用于衡量理论模型与观测数据的契合程度。规范拟合指数（NFI）为0.867，相对拟合指数（RFI）为0.841，略低于建议阈值（≥ 0.90），表明相对于基线模型的比较拟合度仍有小幅提升空间。然而，增值拟合指标表现良好，例如增值拟合指数（IFI = 0.960）、Tucker-Lewis指数（TLI = 0.951）和比较拟合指数（CFI = 0.959），这些指标表明，假设的结构模型相较于基线独立模型有了显著改善。此外，近似均方根误差（RMSEA = 0.050）表明绝对拟合度极佳，明显处于可接受范围内（≤ 0.06），PCLOSE值（0.480）也支持这一结论。总体而言，这些结果表明，结构方程模型（SEM）中提出的结构关系能够有效地反映数据。总的来说，结构方程模型分析显示模型性能良好，其中IFI、TLI、CFI和RMSEA的指标尤为稳健

表 X 模型拟合分析结果。

模型	**NFI (Delta1)**	**RFI (rho1)**	**IFI (Delta2)**	**TLI (rho2)**	**CFI**
默认模型	0.867	0.841	0.96	0.951	0.959
饱和模型	1	1	1	1	1
独立性模型	0	0	0	0	0
模型	**RMSEA**	**90%置信区间下限**	**90%置信区间上限**	**PCLOSE**	
默认模型	0.05	0.031	0.067	0.48	
独立模型	0.226	0.216	0.237	0	

所有假设的中介关系均表现出统计显著性，z 值均超过 1.96。战略在支持与流程之间发挥中介作用，而流程则在战略与生态系统之间发挥中介作用。这些链式中介效应揭示了机器学习项目成功因素的级联特性：组织支持通过战略清晰度、结构化流程和实施质量引发一系列影响，最终决定项目成功。

VI. 讨论

A. 支持 -> 战略（高度显著）

组织支持对战略制定有着显著影响，这与信息系统领域既有文献的观点一致，即组织支持对 IT 项目成功至关重要。组织支持与

战略制定之间的显著关系，在大语言模型辅助开发时代呈现出新的内涵。正如 GitHub Copilot 研究所示，使用人工智能编程助手的开发者完成任务的速度可提高 55%，但要实现这一提速，组织必须制定明确的指导方针和战略框架 [5]。支持维度对于制定人工智能使用政策、伦理准则以及大语言模型工具的资源分配方案至关重要。这与组织支持结构决定人工智能工具究竟能提升生产力还是造成混乱的观点一致 [83]。如今，组织不仅要为传统机器学习项目提供支持，还必须就开发流程中何时以及如何利用大语言模型等更高层面的决策提供支持。

B. 战略 -> 流程（高度显著）

从战略到结构化预处理和后处理的路径揭示了大语言模型辅助开发中的一个悖论。虽然大语言模型能够快速生成预处理代码，但我们的研究结果表明，战略清晰度对于有效实施仍至关重要 [84]。这表明，使用 Copilot 的开发者仍需要清晰地界定问题并理解数据——这些任务无法完全由大语言模型自动完成 [85]。

C. 流程 -> 生态系统（高度显著）

虽然大语言模型擅长生成样板预处理代码，但人工智能生成的代码往往缺乏生产系统所需的健壮性 [86]。我们的研究结果表明，即使利用大语言模型，结构化流程仍然至关重要，这说明大语言模型辅助开发需要新的质量保障和测试方式 [87]。随着团队必须将大语言模型生成的组件与现有基础设施、监控系统和治理框架集成，生态系统视角变得至关重要

D. 生态系统 -> 成功（显著）

生态系统与项目成功之间的关系与近期关于大语言模型辅助开发成果的研究发现一致。尽管大语言模型能够加快个别编码任务，但项目成功取决于系统化的集成实践 [88]。在将模型融入特定组织环境时，生态系统方法至关重要 [89]。成功不仅需要使用这些工具，还需要构建全面的生态系统，以管理版本管理、监控，并持续改进人工编写和人工智能生成的代码。

E. 战略 -> 成功（显著）

战略对成功的直接影响超越了其间接影响，这表明在人工智能增强的环境中，战略清晰度变得愈发重要，而非相反 [90]。直接路径表明，战略除了提升运营效率外，还能创造价值——它影响团队如何构思问题、评估大型语言模型（LLM）的建议，以及做出任何现有人工智能系统都无法完全自动化的架构决策。

F. 流程 -> 成功（显著）

负系数可能代表一种抑制效应，即流程仅通过影响生态系统的实施，才能对成功产生正向作用 [91]。这种统计现象出现在直接效应与总效应符号相反时，表明流程活动会消耗资源并增加复杂性；若缺乏适当的部署基础设施，反而会阻碍成功。

总之，这些发现表明，大语言模型从根本上改变了机器学习项目成功的“如何”，却没有改变其“为何”。虽然编程助手能够加快技术实施，但我们的模型表明，成功仍有赖于组织支持、战略清晰度和结构化流程。这支持“人工智能结对编程”这一范式：大语言模型增强而非取代人类专业知识 [92]。

VII. 局限性

解读结果时，应考虑本研究的若干局限。首先，虽然150名受访者的样本量足以进行结构方程模型分析，但研究结果推广至不同组织情境和行业时可能受到限制。5%的响应率引发了对无回应偏差的担忧，因为完成问卷的受访者可能是具有特定特征或经历的数据科学家。其次，尽管我们的模型有理论依据，横断面设计仍无法进行因果推断；纵向研究更能捕捉机器学习项目演进的动态特性。第三，数据来自自我报告，可能引入共同方法偏差，不过多因子结构和区分效度检验在一定程度上缓解了这一问题。第四，我们聚焦于数据科学家，可能无法代表机器学习项目情境中的所有角色。最后，由于交叉载荷而删除财务支持条目（SP3_FIN），这表明我们对支持维度的操作化可能未能涵盖所有相关组织因素，尤其是与资源分配有关的因素。

VIII. 摘要

本研究通过开发并验证一个综合框架，回应了机器学习项目失败这一长期存在的挑战，并兼顾成功的技术与组织层面。MLC 将商业模式思维应用于机器学习开发这一特定情境。它提供了一种结构化方法，即使在大语言模型编程助手加速开发的情况下依然适用。我们的实证研究结果表明，大语言模型或许能通过快速生成代码，彻底改变机器学习开发的“方式”。然而，根本的成功因素——战略协同、结构化流程、稳健生态系统和组织支持——依然不变。我们的结构方程模型分析揭示了层层递进的关系，可为实践提供参考：组织

不能采用人工智能编码工具并期待成效有所提升。相反，组织必须在全部四个维度上构建全面能力。因此，MLC 框架既可用作诊断工具，帮助识别当前机器学习计划中的薄弱环节，也可作为未来项目的规划工具，帮助组织在现代机器学习开发中驾驭人类专业知识与人工智能辅助的交汇点。

参考文献

[1] A. Ziegler, E. Kalliamvakou, X. A. Li, A. Rice, D. Rifkin, S. Simister, G. Sittampalam 和 E. Aftandilian，“衡量 GitHub Copilot 对生产力的影响”，ACM 通讯，第 67 卷，第 3 期，第 54–63 页，2024年。[2] Anthropic，“软件开发的影响”，Anthropic，2025年4月。[在线]。网址：https://www.anthropic.com/research/impact-software-development [3] T. Weber, M. Brandmaier, A. Schmidt 和 S. Mayer，“通过使用大型语言模型编程显著提升生产力”，ACM 人机交互会议论文集，第 8 卷，第 256 期，第 1-29 页，2024年。[4] L. Gambacorta, H. Qiu, S. Shan 和 D. M. Rees，“生成式人工智能与劳动生产率：一项关于编码的现场实验”，国际清算银行工作论文 1208，2024年。[5] S. Peng, E. Kalliamvakou, P. Cihon 和 M. Demirer，“人工智能对开发人员生产力的影响：来自 GitHub Copilot 的证据”，arXiv 预印本 arXiv:2302.06590，2023年。[6] J. Ryseff, B. F. De Bruhl 和 S. J. Newberry，“人工智能项目失败的根本原因及其成功之道：避免人工智能反模式”，兰德公司研究报告，RR-A2680-1，2024年。[7] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J. F. Crespo 和 D. Dennison，“机器学习系统中隐藏的技术债务”，载于《神经信息处理系统进展》，第 28 卷，第 2503-2511 页，2015年。[8] A. Paleyes, R. G. Urma 和 N. D. Lawrence，“部署机器学习的挑战：案例研究综述”，ACM 计算调查，第 55 卷，第 6 期，第 1-29 页，2022年。[9] D. Kreuzberger, N. Kühl 和 S. Hirschl，“机器学习运维（MLOps）：概述、定义与架构”，IEEE Access，第 11 卷，第 31866-31879 页，2023年。[10] M. Zarour, H. Alzabut 和 K. T. Al-Sarayreh，“机器学习运维最佳实践、挑战与成熟度模型：系统性文献综述”，信息与软件技术，第 183 卷，107733，2025年。[11] M. M. John, H. Holmström Olsson 和 J. Bosch，“机器学习运维的采用实证指南：框架、成熟度模型与分类体系”，信息与软件技术，第 183 卷，107725，2025年。[12] A. Osterwalder，“商业模式本体：一种设计科学方法的提案”，博士论文，洛桑大学，2004年。[13] A. Nurcahyo, F. Rahman 和 T. Wijaya，“人工智能（AI）模型画布框架及其应用案例”，《电气工程、计算机与信息学科学期刊》，第 8 卷，第 1 期，第 1-16 页，2022年。[14] S. Amershi, A. Begel, C. Bird, R. DeLine, H. Gall, E. Kamar, N. Nagappan, B. Nushi 和 T. Zimmermann，“机器学习软件工程：案例研究”，载于《2019 年第 41 届 IEEE/ACM 国际软件工程会议：实践中的软件工程》，第 291-300 页，2019年。[15] M. M. Al-Debei 和 D. E. Avison，“构建统一的商业模式概念框架”，《欧洲信息系统期刊》，第 19 卷，第 3 期，第 359–376 页，2010年。[16] T. Burkhart, J. Krumeich, D. Werth 和 P. Loos，“商业模式概念分析——文献综合分类”，载于《2011 年国际信息系统会议（ICIS）论文集》，2011年。[17] R. Atkinson，“项目管理：成本、时间和质量——两个最佳估计和一种现象；是时候接受其他成功标准了”，《国际项目管理期刊》，第 17 卷，第 6 期，第 337–342 页，1999年。

[18] P. Domingos，《关于机器学习的一些实用知识》，ACM 通讯，第55卷，第10期，第78-87页，2012年。[19] E. Brynjolfsson 和 T. Mitchell，《机器学习能做什么？对劳动力的影响》，《科学》，第358卷，1534，2017 6370，1530 no. pp. - 。[20] C. Rudin，《对于高风险决策，停止解释黑箱机器学习模型，改用可解释模型》，《自然·机器智能》，第1卷，第5期，第206-215页，2019年。[21] J. D. Kelleher、B. Mac Namee 和 A. D'Arcy，《预测性数据分析的机器学习基础》，第2版。美国马萨诸塞州剑桥：麻省理工学院出版社，2020年。[22] N. Mehrabi、F. Morstatter、N. Saxena、K. Lerman 和 A. Galstyan，《机器学习中的偏差与公平性综述》，ACM 计算调查，第54卷，第6期，第1-35页，2021年。[23] N. Sambasivan、S. Kapania、H. Highfill、D. Akrong、P. Paritosh 和 L. M. Aroyo，《人人都想做模型工作，而不是数据工作：高风险人工智能中的数据级联》，载于《2021年CHI计算系统中的人为因素会议论文集》，第1 15页，2021 - 。[24] D. Hendrycks 和 T. Dietterich，《神经网络对常见损坏和扰动的鲁棒性基准测试》，载于《国际学习表征会议论文集》，2019年。[25] B. Goodman 和 S. Flaxman，《欧盟关于算法决策的法规与“解释权”》，《人工智能杂志》，第38卷，第3期，第50-57页，2017年。[26] S. Barocas、M. Hardt 和 A. Narayanan，《公平性与机器学习》。fairmlbook.org，2019年。[27] A. B. Arrieta、N. Díaz-Rodríguez、J. Del Ser、A. Bennetot、S. Tabik、A. Barbado、S. García、S. Gil-López、D. Molina、R. Benjamins、R. Chatila 和 F. Herrera，《可解释人工智能（XAI）：概念、分类体系、机遇与迈向负责任人工智能的挑战》，《信息融合》，第58卷，第82-115页，2020年。[28] P. Chapman、J. Clinton、R. Kerber、T. Khabaza、T. Reinartz、C. Shearer 和 R. Wirth，《CRISP-DM 1.0：数据挖掘分步指南》。SPSS Inc.，2000年。[29] A. Gandomi 和 M. Haider，《超越炒作：大数据概念、方法与分析》，《国际信息管理杂志》，第35卷，第2期，第137-144页，2015年。[30] M. Zaharia、A. Chen、A. Davidson、A. Ghodsi、S. A. Hong、A. Konwinski、S. Murching、T. Nykodym、P. Ogilvie、M. Parkhe、F. Xie 和 C. Zumar，《借助 MLflow 加速机器学习生命周期》，IEEE 数据工程简报，第41卷，45，2018 4，39no. pp. - 。[31] A. Doan、A. Halevy 和 Z. Ives，《数据集成原理》。美国加利福尼亚州旧金山：Morgan Kaufmann，2012年。[32] A. Bifet、R. Gavaldà、G. Holmes 和 B. Pfahringer，《数据流机器学习：附 MOA 实例》。美国马萨诸塞州剑桥：麻省理工学院出版社，2018年。[33] V. Gudivada、A. Apon 和 J. Ding，《大数据与机器学习的数据质量考量：超越数据清洗与转换》，《软件进展国际期刊》，第10卷，第1期，第1-20页，2017年。[34] L. Ehrlinger 和 W. Wöß，《数据质量测量与监控工具综述》，《大数据前沿》，第5卷，850611，2022年。[35] A. Ratner、S. H. Bach、H. Ehrenberg、J. Fries、S. Wu 和 C. Ré，《Snorkel：利用弱监督快速创建训练数据》，《VLDB期刊》，第29卷，第2期，第709-730页，2020年。[36] V. Hodge 和 J. Austin，《异常值检测方法综述》，《人工智能评论》，第22卷，第2期，第85-126页，2004年。[37] A. Zheng 和 A. Casari，《机器学习的特征工程：面向数据科学家的原理与技术》。美国加利福尼亚州塞瓦斯托波尔：O'Reilly Media，2018年。[38] J. Li、K. Cheng、S. Wang、F. Morstatter、R. P. Trevino、J. Tang 和 H. Liu，《特征选择：数据视角》，ACM 计算调查，第50卷，第6期，第1-45页，2017年。[39] D. H. Wolpert 和 W. G. Macready，《优化的无免费午餐定理》，IEEE《进化计算汇刊》，第1卷，第1期，第67-82页，1997年。[40] T. Elsken、J. H. Metzen 和 F. Hutter，《神经架构搜索：综述》，《机器学习研究杂志》，第20卷，第55期，第1-21页，2019年。

[41] F. Zhuang、Z. Qi、K. Duan、D. Xi、Y. Zhu、H. Zhu、H. Xiong 和 Q. He，《迁移学习综合综述》，IEEE会议录，第109卷，第1期，第43-76页，2020年。[42] Z. H. Zhou，《集成方法：基础与算法》。美国佛罗里达州博卡拉顿：CRC出版社，2021年。[43] R. C. Williamson 和 A. K. Menon，《公平性风险度量》，载于《第36届国际机器学习会议论文集》，第6786-6797页，2019年。[44] T. Wu、M. T. Ribeiro、J. Heer 和 D. Weld，《Errudite：可扩展、可复现且可测试的错误分析》，载于《第57届计算语言学协会年会论文集》，第747-763页，2019年。[45] M. Belkin、D. Hsu、S. Ma 和 S. Mandal，《调和现代机器学习实践与经典偏差 - 方差权衡》，美国国家科学院院刊，第116卷，第32期，第15849-15854页，2019年。[46] L. Li、K. Jamieson、A. Rostamizadeh、E. Gonina、M. Hardt、B. Recht 和 A. Talwalkar，《大规模并行超参数调优系统》，机器学习与系统会议录，第2卷，第230-246页，2020年。[47] R. Ashmore、R. Calinescu 和 C. Paterson，《保障机器学习生命周期：期望、方法与挑战》，ACM 计算调查，第54卷，第5期，第1-39页，2022年。[48] H. Suresh 和 J. Guttag，《理解机器学习生命周期中伤害来源的框架》，载于《算法、机制与优化中的公平与可及性》，第1-9页，2021年。[49] T. M. Mitchell，《学习泛化需要偏差》，技术报告 CBM-TR-117，罗格斯大学，1980年。[50] R. Baeza-Yates，《万维网上的偏差》，ACM 通讯，第61卷，第6期，第54-61页，2018年。[51] A. Madry、A. Makelov、L. Schmidt、D. Tsipras 和 A. Vladu，《迈向能够抵御对抗攻击的深度学习模型》，载于《国际学习表征会议》，2018年。[52] R. S. S. Kumar、M. Nyström、J. Lambert、A. Marshall、M. Goertzel、A. Comissoneru、M. Swann 和 S. Xia，《对抗机器学习：业界视角》，载于《2020年IEEE安全与隐私研讨会》，第69-75页，2020年。[53] N. Polyzotis、S. Roy、S. E. Whang 和 M. Zinkevich，《生产环境机器学习中的数据生命周期挑战：综述》，ACM SIGMOD记录，第47卷，第2期，第17-28页，2018年。[54] A. Ratner 等，《MLSys：机器学习系统的新前沿》，arXiv预印本 arXiv:1904.03257，2019年。[55] T. Li、A. K. Sahu、A. Talwalkar 和 V. Smith，《联邦学习：挑战、方法与未来方向》，IEEE信号处理杂志，第37卷，第3期，第50-60页，2020年。[56] E. Jonas、J. Schleier-Smith、V. Sreekanti、C. C. Tsai、A. Khandelwal、Q. Pu、V. Shankar、J. Carreira、K. Krauth、N. Yadwadkar、J. E. Gonzalez、R. A. Popa、I. Stoica 和 D. A. Patterson，《简化云编程：伯克利对无服务器计算的看法》，arXiv预印本 arXiv:1902.03383，2019年。[57] M. Satyanarayanan，《边缘计算的兴起》，Computer，第50卷，第1期，第30-39页，2017年。[58] R. Bhardwaj、A. Nambiar 和 D. Dutta，《云计算中的机器学习研究》，IEEE云计算汇刊，第10卷，第1期，第166-183页，2022年。[59] S. Raschka、J. Patterson 和 C. Nolet，《Python 中的机器学习：数据科学、机器学习与人工智能的主要发展和技术趋势》，Information，第11卷，4，193，2020 no. 。[60] M. Zaharia、R. S. Xin、P. Wendell、T. Das、M. Armbrust、A. Dave、X. Meng、J. Rosen、S. Venkataraman、M. J. Franklin、A. Ghodsi、J. Gonzalez、S. Shenker 和 I. Stoica，《Apache Spark：用于大数据处理的统一引擎》，ACM 通讯，第59卷，第11期，第56-65页，2016年。[61] A. Rule、A. Tabard 和 J. D. Hollan，《计算笔记本中的探索与解释》，载于《2018年CHI计算系统中的人为因素会议论文集》，第1 12页，2018 - 。[62] E. Breck、N. Polyzotis、S. Roy、S. E. Whang 和 M. Zinkevich，《机器学习的数据验证》，机器学习与系统会议录，第1卷，第334-347页，2019年。[63] D. Kang、D. Raghavan、P. Bailis 和 M. Zaharia，《用于监控和改进机器学习模型的模型断言》，机器学习与系统会议录，第2卷，第481-496页，2020年。

[64] T. Ben-Nun and T. Hoefler，《揭开并行与分布式深度学习的神秘面纱：深入的并发分析》，ACM 计算调查，第52卷，第4期，第1-43页，2019年。[65] J. Gama、I. Žliobaitė、A. Bifet、M. Pechenizkiy和A. Bouchachia，《概念漂移适应综述》，ACM 计算调查，第46卷，第4期，第1-37页，2014年。[66] D. Xin、L. Ma、J. Liu、S. Macke、S. Song和A. Parameswaran，《加速人在回路中的机器学习：挑战与机遇》，载于《第四届端到端机器学习数据管理研讨会论文集》，第1-8页，2021年。[67] S. Ransbotham、S. Khodabandeh、R. Fehling、B. LaFountain和D. Kiron，《凭借人工智能取胜》，《麻省理工斯隆管理评论》，第61卷，第1期，第17-19页，2019年。[68] VentureBeat，《为什么87%的数据科学项目从未投入生产？》，2019年7月。[在线]。https://venturebeat.com/ai/why-do-87-of-data-science-projects-never-make-it-into-production/2019。[69] T. H. Davenport和D. J. Patil，《数据科学家：21世纪最性感的职业》，《哈佛商业评论》，第90卷，第10期，第70-76页，2012年。[70] Google Cloud，《MLOps实践者指南：机器学习持续交付与自动化框架》，Google Cloud架构框架，2021年。[71] E. Glikson和A. W. Woolley，《人类对人工智能的信任：实证研究综述》，《管理学会年刊》，第14卷，第2期，第627-660页，2020年。[72] A. Ng，《人工智能转型实战指南》，Landing AI，2021年。[73] J. K. U. Brock和F. von Wangenheim，《揭开人工智能的神秘面纱：数字化转型领导者能教会你如何务实地应用人工智能》，《加州管理评论》，第61卷，第4期，第110-134页，2019年。[74] P. Saleiro、B. Kuester、L. Hinkson、J. London、A. Stevens、A. Anisfeld、K. T. Rodolfa和R. Ghani，《Aequitas：偏见与公平性审计工具包》，arXiv预印本 arXiv:1811.05577，2018年。[75] 欧盟委员会，《关于制定人工智能协调规则的条例提案（《人工智能法案》）》，COM(2021) 206 final，2021年。[76] J. G. March，《组织学习中的探索与利用》，《组织科学》，第2卷，第1期，第71-87页，1991年。[77] J. Barney，《企业资源与持续竞争优势》，《管理学杂志》，第17卷，第1期，第99-120页，1991年。[78] E. Brynjolfsson和K. McElheran，《数据驱动决策的迅速普及》，《美国经济评论》，第106卷，第5期，第133-139页，2016年。[79] T. H. Davenport和J. G. Harris，《以分析取胜：制胜新科学》。美国马萨诸塞州波士顿：哈佛商业出版社，2007年。[80] J. C. Henderson和N. Venkatraman，《战略协同：利用信息技术推动组织转型》，《IBM系统杂志》，第32卷，第1期，第4-16页，1993年。[81] T. Fountaine、B. McCarthy和T. Saleh，《构建人工智能驱动型组织》，《哈佛商业评论》，2019年7月至8月。[82] K. Wagstaff，《有意义的机器学习》，载于《第29届国际机器学习会议论文集》，第536-529页，2012年。[83] E. Brynjolfsson、D. Li和L. Raymond，《工作中的生成式人工智能》，美国国家经济研究局工作论文，第w31161号，2023年。[84] M. Chen等，《评估基于代码训练的大型语言模型》，arXiv预印本 arXiv:2107.03374，2021年。[85] A. Sarkar、A. D. Gordon、C. Negreanu、C. Poelitz、S. S. Ragavan和B. Zorn，《使用人工智能编程是什么体验？》，arXiv预印本 arXiv:2208.06213，2021年。[86] J. D. Weisz、M. Muller、S. Houde、J. Richards、S. I. Ross、F. Martinez、M. Agarwal和K. Talamadupula，《不必追求完美？代码转换中的人机协作》，载于《第26届智能用户界面国际会议》，第402-412页，2021年。[87] T. Zhang、V. Kishore、F. Wu、K. Q. Weinberger和Y. Artzi，《BERTScore：使用BERT评估文本生成》，arXiv预印本 arXiv:1904.09675，2023年。[88] P. Vaithilingam、T. Zhang和E. L. Glassman，《期望与体验：评估大型语言模型驱动的代码生成工具的可用性》，载于《CHI计算机系统人为因素会议扩展摘要》，第1-7页，2022年。

[89] R. Bommasani等，《基础模型的机遇与风险》，arXiv预印本arXiv:2108.07258，2021年。[90] A. Kaplan和M. Haenlein，《Siri，Siri，在我手中：谁是世间最公平的？论人工智能的解读、阐释与影响》，*Business Horizons*，第62卷，第25期，第1-15页，2019年。[91] D. P. MacKinnon、J. L. Krull和C. M. Lockwood，《中介效应、混杂效应与抑制效应的等价性》，*Prevention Science*，第1卷，第4期，第173-181页，2000年。[92] S. Barke、M. B. James和N. Polikarpova，《有依据的编程助手：程序员如何与代码生成模型交互》，*ACM* 编程语言会议论文集，第7卷，OOPSLA1期，第85-111页，2023年。

Martin Prause：作为一名科学家和企业家，他结合经济学、计算机科学和行为经济学的方法，使个人和组织层面的决策流程变得透明、易于理解。他是一家公司的联合创始人兼首席技术官，该公司通过模拟展现企业决策，以优化决策，并将由此获得的洞见成功融入商业战略。