迈向端到端机器学习运维工具图谱：基于多声部文献综述的初步研究

Sergio Moreschini[a], Gilberto Recupito[b,a], ValentinaLenarduzzi[c], FabioPalomba[b], DavidH¨astbacka[a], Davide Taibi[c,a]

[a] *Tampere University, Finland*[b] *Software Engineering (SeSa) Lab, Department of Computer Science, University of Salerno, Italy*[c] *University of Oulu, Finland*

arXiv:2304.03254v1 [cs.SE] 6 Apr 2023

摘要

机器学习运维工具遵循开发运维流程，实现机器学习的持续开发。市场上已经推出了不同的机器学习运维工具，然而，如此多的工具往往会导致人们对每个开发运维阶段应使用哪种最合适的工具感到困惑。为解决这一问题，我们开展了一项多声部文献综述，对从254项主要研究中识别出的84种机器学习运维工具在开发运维阶段中的分布进行了映射，重点介绍了它们的用途以及可能存在的不兼容性。这项工作的成果将有助于实践者和研究人员，并可作为未来研究机器学习运维工具、流水线和流程的起点。

关键词： 机器学习运维、人工智能运维、开发运维、人工智能、机器学习

1. 引言

机器学习运维（MLOps）是一种实践，它结合软件工程和数据科学的最佳实践，用于管理机器学习模型的端到端生命周期，并将其集成到传统软件中。机器学习运维涉及使用工具和技术，自动化机器学习模型的构建、测试、部署和监控。其目标是在降低错误风险的同时，提高机器学习模型的开发速度、质量和可靠性。

随着部署到生产环境中的机器学习模型数量持续增长，机器学习运维工具正变得越来越重要。这些工具有助于数据科学家和工程师简化机器学习模型的构建、测试和部署流程，并使生产环境中的模型更易于监控和维护。这使组织能够更灵活地运用机器学习，并快速适应不断变化的业务需求和用户反馈。

市场上已经推出了不同的机器学习运维工具 [1]，，大量从业者会议开始介绍机器学习运维及相关工具。

然而，机器学习运维工具数量极其庞大，这可能会给需要从中选择一部分来创建流水线的组织带来问题。此外，并非所有的工具可以组合在一起，在采用这些工具的过程中引入其他问题。

在我们之前的工作中 [1], ，我们提出了一项多声部文献综述，以研究Google搜索前102 条结果以及 Google 学术中提及的机器学习运维工具的特征。然而，我们指出，之前的工作存在若干局限性，包括搜索范围有限，以及未对所选来源进行滚雪球检索。

本研究的最终目标是为研究人员和从业者提供一些初步见解，帮助他们了解端到端机器学习流水线场景下可用的工具。一方面，研究结果或许能帮助研究人员了解当前实际环境中为机器学习工程师提供的支持。另一方面，我们的发现或可为后续研究奠定基础，以了解从业者会在实践中何处使用现有工具，以及他们会如何构建理想的流水线，从而更有效地支持端到端机器学习。

因此，我们对先前工作进行了一次内部差异化复现 [2] ，克服其局限，并进一步识别各工具可应用于哪些不同的开发运维阶段，同时确定工具之间的兼容性与不兼容性。本研究的主要贡献是绘制了一幅机器学习运维工具的图形化开发运维图，帮助从业者和研究人员轻松识别适用于各个开发运维阶段的工具。

本文其余部分的结构如下。第 2 节介绍了开发运维和机器学习运维的背景。第3 节描述了复现研究 [1] 并强调了其与我们工作的差异。第 4 节介绍

Email addresses: sergio.moreschini@tuni.fi(Sergio Moreschini), grecupito@unisa.it (Gilberto Recupito),valentina.lenarduzzi@oulu.fi (Valentina Lenarduzzi),fpalomba@unisa.it(Fabio Palomba), david.hastbacka@tuni.fi(David H¨astbacka), davide.taibi@oulu.fi (Davide Taibi)

本研究所针对的研究问题以及为解决这些问题所采用的方法。第 5 节介绍研究结果，第 6 节进一步阐述这些结果。第 7 节指出并报告了本研究有效性面临的威胁。第 8 节讨论了机器学习运维领域的相关工作，并重点说明本研究如何推动该领域的发展。最后，第 9 节总结全文并指出未来的研究方向。

2. 背景

开发运维是一套旨在改善开发团队与运维团队之间协作和沟通的实践与工具，以加快软件交付流程。开发运维的目标是在提高软件发布速度和质量的同时，降低故障和停机风险 [3]。

开发运维流程通常包括以下阶段：

- **计划**。 代码开发之前开展的所有活动。识别业务需求并收集最终用户反馈。目标是为未来的开发制定产品路线图。
- **编码**。所有严格与编码开发阶段相关的活动。此阶段使用的工具包括用于辅助开发流程的插件。
- **构建**。编码任务完成后，开发人员会将代码提交到共享代码仓库以生成构建。
- **测试**。 构建将在测试环境中进行部署，以接受多种类型的测试。这些测试包括验收测试、安全测试、集成测试、性能测试等。
- **发布**。 构建成功通过所有自动化测试和手动测试后，即可部署到生产环境，运维团队可以安排发布。
- **部署**。 构建已准备就绪，并将其发布到生产环境中。一组工具用于自动化发布流程。
- **操作**。 在此阶段，发布已完成，并已被客户使用。运维团队负责服务器的配置和供应。
- **监控**。 根据在操作阶段获取的数据，对整个流水线进行监控。

机器学习运维（MLOps）是一套旨在改善数据科学家与IT/DevOps团队在构建、测试和部署机器学习模型时协作与沟通的实践和工具。它旨在帮助组织管理部署机器学习模型的复杂性在生产环境中，并确保模型
可靠、可扩展且易于维护。

MLOps 与传统 DevOps 的一个关键区别在于，MLOps 专门关注机器学习模型的需求，而这些需求可能与传统软件大不相同。例如，机器学习模型需要更多数据进行训练，而且通常比传统软件需要更高的计算量。此外，机器学习模型还容易受到不同类型错误的影响，例如过拟合和偏差，而这些错误可能难以检测和修复。

另一个关键区别是，机器学习运维通常涉及更复杂、数据密集型的工作流。数据科学家可能需要处理大量数据、执行特征工程，并试验不同的模型架构，而所有这些工作都可能耗时且资源密集。机器学习运维通过自动化构建、测试和部署机器学习模型过程中涉及的许多任务，帮助弥合数据科学家与 IT/开发运维团队之间的差距，例如数据预处理、模型训练和模型服务。

机器学习运维还有助于模型治理与管理。机器学习模型需要部署、监控和维护，而机器学习运维提供了自动化完成这些任务所需的基础设施和工具。它还有助于跟踪模型的版本和性能指标，并提高用于构建模型的数据的透明度。

总之，机器学习运维是一套实践和工具，可帮助组织应对在生产环境中部署机器学习模型的复杂性。它聚焦于机器学习模型的特定需求，并通过自动化构建、测试和部署机器学习模型所涉及的诸多任务，帮助弥合数据科学家与 IT/开发运维团队之间的鸿沟。它还有助于模型治理和管理，并提供自动化执行这些任务所需的基础设施和工具。

Moreschini 等人提出的机器学习运维流程图示 [4] 如（图1）所示。

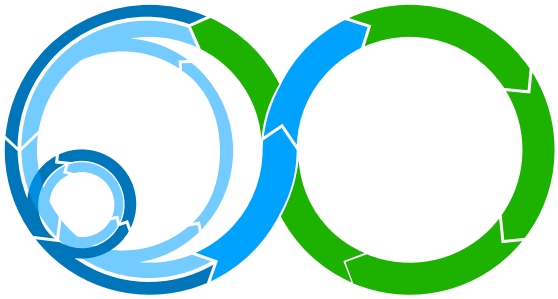

图1：机器学习运维无限循环 [4]

表1：研究比较

	Recupito 等人 [1]	**我们的研究**
RQs	- 机器学习运维工具 - 机器学习运维工具的主要功能	- 将机器学习运维工具映射到机器学习运维阶段
检索式	我们在检索式中增加了更多检索词，包括”机器学习运维”和”机器学习运维”	
参考文献	Google 学术	Google 学术、ACM数字图书馆、IEEEX-探索数字图书馆、Science Direct、Scopus 和 Springer Link
灰色文献搜索引擎	Google	Google、Twitter、Reddit 和 Medium
滚雪球检索	未执行	已执行
纳入标准（主要研究）	- 该研究讨论了最小化组件 端到端机器学习运维工作流 - 本研究讨论机器学习运维实践或 基于机器学习的应用程序 - 本研究指的是.....的实施。 机器学习运维工具 - 该研究讨论经验、观点或 关于机器学习运维流水线的实践	- 提及可用于创建的工具的主要研究 或编排 MLOps 流水线
排除标准（主要研究）	- 该研究未提供有关.....的详细信息 MLOps 工具的设计或实现 - 该研究仅提供特定对象的设计 机器学习流水线的组成部分 - 该研究未提供或提及相关细节- 关于机器学习自动化 - 本研究涉及商业平台，其 提供 MLOps 应用以销售其服务 用于开发和部署	- 研究计划、路线图、愿景论文、Mas- 硕士论文、书籍 - 作者的关注者少于 750 名关注者（仅适用于 Medium - 灰色文献)
纳入标准（工具）	从数据提取到模型监控工具	端到端 机器学习运维工具, 包括 所有 机器学习运维流程的各个阶段（图1）
排除标准（工具）	1）该机器学习运维工具未提供文档 实施或未列出特征 - 提及机器学习运维工具的文献少于五篇 所选研究的次数	- 非必需项（训练编排、Explain- 可实现的 AI, 特征存储) - 阻碍自动化的工具 eration, 数据标注, 数据监控) - 附加组件 - 不可靠（例如，星级不足）
工具分类	- 主要 MLOps 特征 - 与机器学习相关的数据管理功能 模型 - 机器学习模型管理功能	- 每个工具涵盖的机器学习运维流水线步骤 - 之间的兼容性与不兼容性 工具 - 机器学习运维工具的可能组合 流水线
选定的主要研究	60（6 个白色，54 个灰色）	254（51 个白色，203 个灰色）
机器学习运维工具	13	84

3. 复现研究

本节将概述复现研究 [1]以及我们复现该研究的原因，并着重说明其与我们工作的差异。

我们决定复现该研究 [1], ，因为——据我们所知——这是唯一一项识别出一组机器学习运维工具及其特征的二手研究。

本次复现遵循 Carver 等人提出的指南 [2]。

Recupito 等人 [1] 开展了一项多声部文献综述，研究了可应用于机器学习流水线的机器学习运维工具，涵盖从数据提取到模型监控的各个阶段。他们考察了灰色文献和白色文献来源，并列出了这些工具及其主要差异特征，重点关注：1）机器学习运维的主要特征；2）与机器学习模型相关的数据管理特征；3）机器学习模型管理特征。此外，他们将这些工具分为基于云的机器学习平台、编排平台，或 TensorFlow Extended。他们从 Google 学术中提取白色文献，并将 Google 搜索的前 102 条结果作为灰色文献来源。在这 102 条结果中，他们找到 96 个网站（包括博客文章和开发者网站）、9 个 GitHub 仓库和 5 个 YouTube 视频。应用纳入和排除标准后，他们共纳入 60 个来源，其中包括 42 个网站/博客、7 个 GitHub 项目、5 个 YouTube 视频，以及 6 篇来自 Google 学术的同行评审文献。

因此，他们从选定的 60 个来源中提取了 13 种工具。

3.1. 与本研究的差异

与我们的原始研究 [1], 不同，我们考虑了可应用于机器学习运维流水线所有步骤的工具，扩展了检索式，并为白色文献采用了六个书目来源而非一个，以及为灰色文献采用了三个搜索引擎而非一个。

此外，我们进行了滚雪球过程，并放宽了纳入和排除标准，以尽可能获取更多相关论文。因此，我们纳入了254个来源，从中提取工具，而 [1], 仅考虑了60个来源，并获得了包含84个工具的列表，而不是13个。

除了我们分类的工具数量更多之外，最重要的结果在于复现研究的关注点有所不同。在此前的工作 [1]中，我们识别了这些工具的主要特性；而本研究聚焦于工具可应用的机器学习运维阶段，以及创建工具流水线时它们之间的兼容性与不兼容性。

本研究与复现研究 [1] 的完整比较见表 1。

4. 研究设计

我们的研究目标是绘制所有机器学习运维工具的图谱，旨在概述一条获得全面支持的机器学习运维流水线。因此，为了探究上述目标，我们提出了以下研究问题（RQ）：

研究问题 工具在多大程度上支持机器学习运维的不同阶段?

我们旨在了解机器学习运维流程的不同阶段可以使用哪些工具。

为了以客观、系统且可复现的方式回答这一研究问题，我们采用了系统性多元文献综述（MLR）方法，并遵循 Garousi 等人提出的指南 [5]。

多元文献综述流程既包括同行评审文献，也包括灰色文献，并且在结果中考虑了从业者与学术研究人员之间的不同观点。多元文献综述流程将同行评审论文的贡献归类为学术文献，将博客文章、白皮书、播客等其他类型内容的贡献归类为灰色文献。

多元文献综述流程由四个步骤组成，如图2所示：

- 灰色文献和白色文献中的主要研究的筛选；
- 质量评估 所选灰色文献；
- 数据提取 回答我们研究问题所需的信息；
  - 从主要研究中提取数据– 从工具网站提取数据
- 工具选择
- 数据综合 提取结果

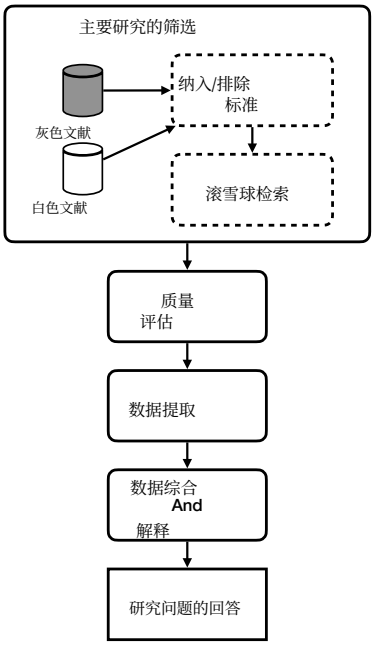

图2：所采用的多元文献综述流程概览

4.1. 主要研究的筛选

筛选主要研究（PS）的第一步是确定检索式，以便在学术书目数据库和灰色文献搜索引擎中使用。

为提高查全率并尽可能纳入更多论文，我们扩展了 Recupito 等人 [1] 使用的检索式：

(“机器学习运维” OR “机器学习运维” OR 机器学习运维 OR “机器学习运维” OR “机器学习运维” OR 数据运维) AND (工具 OR 应用程序) AND (生命周期 OR 流水线 OR 平台 OR 工作流)

搜索词应用于所有字段（即标题、摘要和关键词），以尽可能纳入更多研究成果。对于从在线来源检索灰色文献以及从学术文献数据库检索白色文献，我们采用了相同的搜索词。

同行评审文献检索。我们考虑了六个书目来源所收录的论文：

- ACM数字图书馆[1]
- IEEEXplore 数字图书馆[2]
- Science Direct[3]

[1] https://dl.acm.org
[2] https://ieeexplore.ieee.org
[3] https://www.sciencedirect.com

- Scopus[4]
- Google 学术[5]
- Springer Link[6]

两次检索（灰色文献和白色文献）均于2022年11月进行。

灰色文献检索。我们使用四个搜索引擎进行了检索：

- Google 搜索[7]
- Twitter[8]
- Reddit[9]
- Medium[10]

搜索结果包括书籍、博客文章、论坛、网站、视频、白皮书、框架和播客。

滚雪球检索。滚雪球检索是指利用论文的参考文献列表或引用该论文的文献来识别其他主要研究 [6]。我们对学术文献应用了反向滚雪球检索，通过所选来源的参考文献来识别相关的主要研究。此外，我们还对灰色文献应用了反向滚雪球检索，沿着每个所选来源的外部链接进行检索。

应用程序纳入和排除标准。基于系统性文献综述指南 [7], 我们定义了纳入和排除 标准（表 2 ）。我们采用了限制性较低的纳入标准，以便纳入更全面的工具集。

在应用纳入和排除标准之前，我们在从检索结果中随机选取的25项主要研究子集上测试了这些标准的适用性 [8] 。最终的纳入和排除标准汇总于表2中。

为筛选每篇论文，我们安排两名研究人员独立审阅。为确保公平，我们打乱了任务分配，并确保每位研究人员负责审阅的论文数量与团队其他成员大致相同。如有分歧，则请第三位作者参与，以达成共识。此举旨在提高研究的可靠性。最后，为评估引入第三位作者前的评审者间一致性，我们计算了 Cohen's kappa 系数。相关结果记录在复现包 [11] 中。

[4] https://www.scopus.com
[5] https://scholar.google.com
[6] https://link.springer.com/
[7] https://www.google.com/
[8] https://twitter.com/
[9] https://www.reddit.com/
[10] https://medium.com

4.2. 灰色文献的质量评估

与经过同行评审的文献不同，灰色文献未经正式评审，因此其质量较难得到有效控制。为评估所选灰色文献来源的可信度和质量，并决定是否纳入某一灰色文献来源，我们遵循了 Garousi 等人的指南建议 [5], ，考量了发布者的权威性、所采用的方法、客观性、日期、新颖性和影响力。

前两位作者根据上述标准对每个来源进行评估，并根据标准本身采用二元或三点李克特量表。如有分歧，我们会与第三位作者讨论评估结果，由其协助给出最终评估。

4.3. 数据提取

由于我们的目标是描述机器学习运维工具中的信息，因此我们需要直接从工具网站获取这些信息。因此，数据提取过程由两个步骤组成：

(PE) 从符合质量评估标准的主要研究（*PSs*）中提取工具列表。(TE) 从工具列表中提取信息。在这种情况下，我们直接从官方网站门户提取信息。

基于RQ，我们在审阅电子表格中提取了信息。数据提取表以及为回答RQ所需信息的映射总结于表3中。

数据提取过程遵循Wohlin 等人提出的定性分析指南 [9]。两名研究人员提取了所有信息。如有分歧，我们会邀请第三位作者参与讨论。讨论一直持续到分歧得到解决为止。

4.4. 工具选择

根据上一步骤中提取的工具列表，我们需要确定最终工具集，以回答我们的 RQ。因此，我们采用了与论文选择阶段（第 4.1 节）类似的流程，基于一组纳入和排除标准对工具进行筛选。

在应用纳入和排除标准之前，我们在从检索到的工具中随机选取的 10% 子集上测试了这些标准的适用性 [8]。最终的纳入和排除标准集见表 2。

与主要研究的选择一样，工具的选择也由两名研究人员完成，他们分别对其进行了审阅。同样在这种情况下，如有分歧，则会邀请第三位作者参与，以达成共识。在此情况下，也计算了 Cohen's kappa 系数，以确认评审者之间的一致性。这些结果记录在复现包 [11] 中。

表 2：纳入/排除标准

	标准
纳入	提及可用于创建或组合机器学习运维流水线的工具的主要研究
排除	- 非英文 - 重复内容（总结其他网站内容的帖子） - 主题不相关（将这些术语用于其他用途） - 未经同行评审的论文 - 研究计划、路线图、愿景论文、硕士论文 - 作者的关注者少于 750 名（仅适用于 Medium 上的灰色文献）*
工具选择	**标准**
纳入	商业和开源工具
排除	- 无法下载 - 无可用/可访问的网站 - 非英文 - 重复 - 硬件 - 不可靠（例如：GitHub 星标数量不足） - 非必需项（例如：训练编排、可解释人工智能、特征存储） - 附加组件 - 工具破坏自动化（数据生成、数据标注）

* 我们设定了这一阈值　　因为某些账号仅为推广工具而开设，或者作者没有获得足够的关注者，因　　e
与他们之前的出版物相比　ons

表3：数据提取

Info	**Description**	**Step**
工具名称 工具网址	工具名称	PE
阶段	所覆盖的主要开发运维阶段为该工具（计划，　代码，　构建，测试、发布、部署、操作，监控）	TE
用途	工具的主要用途（例如：持续训练）	
替代方案	该工具的主要竞争对手	

4.5. 开展综述

通过检索过程，我们共检索到 783 篇不重复的主要研究（排除 137 篇重复文献后）：其中 497 篇来自灰色文献，287 篇来自白色文献。滚雪球过程又纳入了 30 篇文献（其中 26 篇来自白色文献，4 篇来自灰色文献）。

应用纳入和排除标准后，我们选取了 254 篇主要研究（其中 51 篇来自白色文献，203 篇来自灰色文献）。作者应用纳入标准所得结果的一致性几乎完美（Cohen's kappa $= 0.870$）。

质量评估流程应用于灰色文献主要研究，导致排除 1 项主要研究。

因此，我们共纳入254项主要研究，其中203项来自灰色文献，51项来自白色文献。

通过对选定的254项主要研究进行数据提取，我们获得了84种机器学习运维工具。

应用纳入和排除标准后，结果显示几乎完全一致（Cohen's Kappa 系数为 $= 0.812$）。因此，我们最终得到 84 篇文献，如表 ?? 所示。

4.6. 可验证性与可复现性

为便于复现我们的研究，我们已在复现包中发布完整的原始数据。[11].

5. 结果分析

我们提取了已识别工具及其替代方案的特征，并根据各工具所覆盖的机器学习运维阶段，将其归入以下子类别。我们确定了5个全局类别：

1. **持续开发（持续交付）**：根据开发运维的经典定义，持续交付包括能够创建单个软件包或服务的工具。为将这一概念扩展到机器学习运维，我们还纳入了能够创建自动部署另一项服务的系统的工具 [10]。2. **持续集成（持续集成）**：根据开发运维的经典定义，持续集成包括用于测试和验证代码及组件的工具。为将这一概念扩展到机器学习运维，我们还纳入了用于测试和验证数据及模型的工具 [10]。3. **持续测试（持续训练）**：用于自动重新训练模型并提供模型服务的工具 [10]。
4. **端到端全栈机器学习运维工具**：包括可单独用于创建完整流水线的工具。

5. **运维（OPS）**：根据开发运维的经典定义，运维包括用于管理发布、部署、操作和监控阶段的工具。

[11] https://figshare.com/s/0b62981a4ed90a93099f

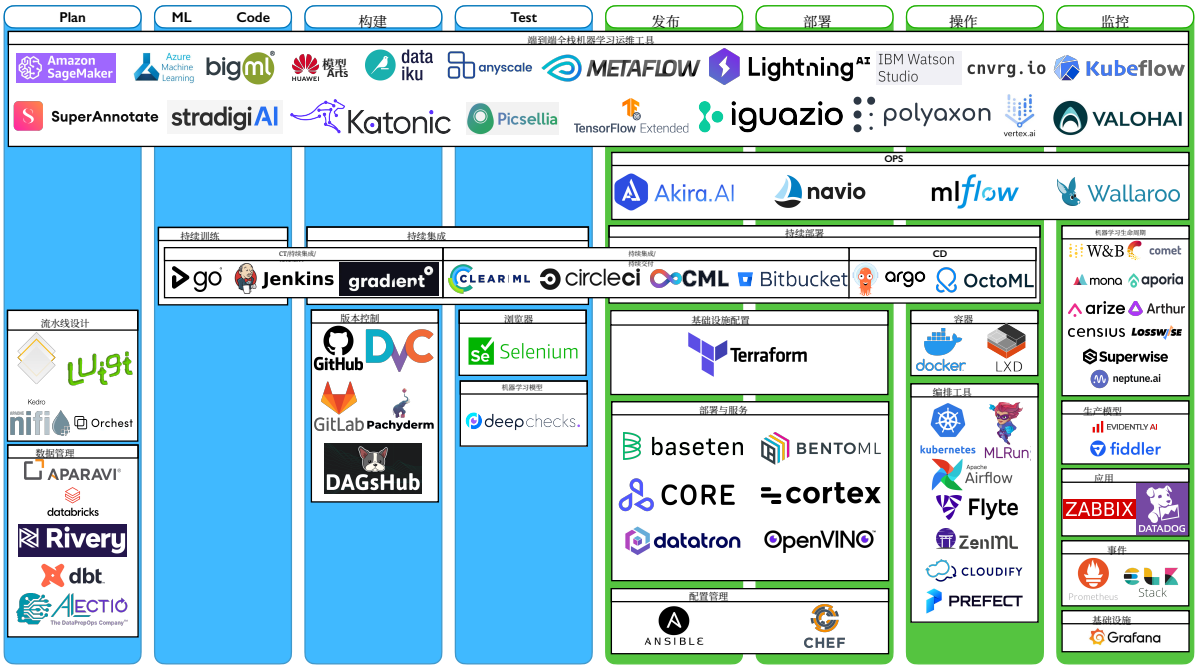

图3：用于构建机器学习运维流水线的已识别工具及其子类别

随后，我们结合第 2 节所述的各个阶段，进一步完善了分类。

计划。我们确定了 2 个子类别：

1. 数据管理：负责管理用于训练机器学习模型的数据的工具。 2.流水线设计：用于创建流水线的工具。这些工具可以包括依赖解析、工作流管理和可视化。

构建。作为子类别，我们确定了版本控制：负责跟踪和管理代码变更的工具

测试。我们确定了 2 个子类别：

1. 代码：基于代码的测试工具。2. 机器学习模型：用于测试机器学习模型及其训练数据的工具。

发布/部署。我们确定了 3 个子类别：

1. 配置管理：用于执行配置管理的工具，即在产品的整个生命周期内保持其属性的一致性。2. 部署与服务：用于自动将机器学习模型部署到生产环境中的工具。此过程会自动完成许多子任务，包括模型服务，这是一种将机器学习模型集成到软件系统中的技术。

3. 基础设施配置：用于配置或创建基础设施资源的工具。

操作。我们确定了 2 个子类别：

1. 容器：用于创建容器的工具。容器被定义为“将代码及其所有依赖项打包在一起的标准软件单元，使应用程序能够在不同的计算环境之间快速、可靠地运行” [11]。

2. 编排工具：用于自动化容器协调流程的工具。包括启动和停止、任务调度与执行，以及恢复流程。

监控。我们确定了 5 个子类别：

1. 应用：用于检查和分析已创建应用程序性能数据的工具。2. 事件：用于监控事件并提供警报功能的工具。3. 基础设施：用于监控基础设施节点的工具。4. 机器学习生命周期：用于跟踪实验、对数据集进行版本管理，以及管理机器学习模型的工具。5. 生产环境中的模型：用于监控已部署的机器学习模型，以避免模型在生产环境中发生故障或性能下降的工具。

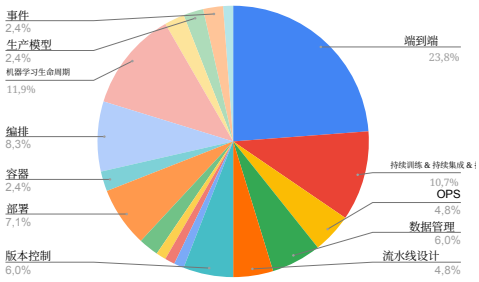

图4：按子类别划分的选定工具

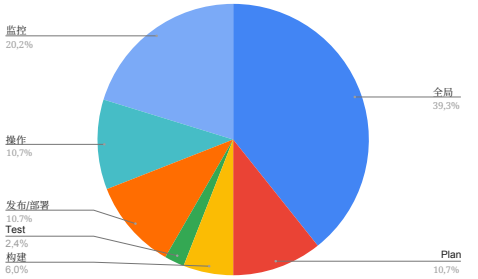

图 5：各阶段的选定工具

不同机器学习运维工具的分类最终结果如图3所示。共有84种不同工具，被划分为20个不同类别。各子类别包含的不同工具数量见图4和表4，其中持续训练、持续集成和持续交付工具已归为一组。

图4展示了根据工具所支持的机器学习运维流水线阶段划分的工具分布。此外，在本例中，持续训练、持续集成和持续交付已被归为一组。

表4：工具和类别

类别	子类别	# 工具
全局	- 持续训练 & 持续集成 & 持续交付	9
	- 端到端全栈机器学习运维工具	20
	- 运维	4
Plan	数据管理	5
	- 流水线设计	4
构建	版本控制	5
Test	- 浏览器	1
	机器学习模型	1
发布 & 部署	- 部署与服务	6
	- 基础设施配置	1
操作	- 配置管理	2
	- 容器	2
	- 编排工具	7
监控	- 应用	2
	- 事件	2
	- 基础设施	1
	- 机器学习生命周期	10
	- 生产模型	2

6. 讨论与启示

研究结果提供了若干值得进一步讨论的启示和可行动要点。

关于机器学习运维工具的存在。

首先，我们的研究结果揭示了84种机器学习运维工具的存在。如此大量的工具主要由工具供应商和从业者提供——灰色文献综述分析中所识别出的工具数量印证了这一点。一方面，这凸显了从业者对机器学习运维活动日益增长的关注，进一步推动了围绕这一主题的研究，也凸显了有必要全面分析如何组合这些工具，而这正是我们研究初步探讨的内容。另一方面，我们指出，研究界提出或通过实证研究考察的工具明显不足。特别是，目前可用机器学习运维工具的完整清单，以及将这些工具映射到机器学习运维流水线，均是我们研究的主要成果，或可作为从不同角度进一步研究该问题的基础，例如这些工具在实践中的可操作性、从业者有效使用这些工具所需的信息，或组合多个工具的最有效策略。

启示 *#1*。因此，我们的研究或许是提高人们对实证软件工程研究领域可能对新型机器学习运维解决方案的开发产生何种影响的认识的第一步。

关于机器学习运维工具的兼容性。

“端到端机器学习运维工具”通常会提供预定义的工具流水线。不过，在某些情况下，可以用其他工具替换其中的工具。这类工具需要覆盖机器学习运维流水线的所有阶段。因此，若要实现最佳流水线生成而不采用“端到端机器学习运维工具”，就需要为“开发”阶段的每个阶段使用一个工具，并为“运维”阶段至少使用一个工具[12]。这是因为，虽然在“运维”阶段纳入多个工具可能会产生有价值且经过改进的输出，但在“开发”的同一阶段纳入多个工具会造成冗余，并可能破坏流水线自动化。根据我们的映射工作，系统性研究中涌现出的工具似乎彼此兼容，也就是说，它们能在整个机器学习运维流水线中提供相互正交的信息。

启示*#2*。我们认为，这一结果可能促成未来针对该问题的若干研究：涉及所需信息流的问题，旨在

依次使用工具，或利用多个工具可能为实践者提供的冗余/对比信息，这些问题对于我们的研究群体而言确实仍是悬而未决且被忽视的。

关于分类体系的创建。
最后但同样重要的是，我们开展的映射工作可能会对工具供应商和从业者产生影响。前者可能会发现一份关于当前技术/实践水平的全面概览，这份概览有望作为开发新型工具的基础，用于覆盖机器学习运维流水线中目前代表性较低的步骤。此外，这份概览还可能促成进一步的思考，例如对这些工具所提供功能进行基于充分信息的分析，进而促使工具供应商改进现有工具和/或设计额外的支持系统。

影响 *#3*。从业者可以利用我们的分类体系，通过发现此前不了解的替代工具，或尝试组合使用能够帮助其覆盖整个机器学习运维流水线的工具，来调整自身的 *MLOps* 流水线。

7. 有效性威胁

我们的研究结果可能受到各种偏差或误差来源的影响，包括数据提取不准确、文献综述范围的局限性，以及纳入标准和排除标准的定义与应用具有主观性。在本节中，我们将根据 [9] 中所述的指南，概述为降低这些潜在威胁而采取的策略。

构念效度。根据参考文献，构念效度关注的是研究对象在多大程度上准确反映了研究背后的理论。

研究问题和研究中使用的分类方案可能受到此类威胁的影响。为尽量降低这一风险，作者分别审阅了研究问题，随后进行了讨论。至于采用的分类方案，我们选择了开发运维的标准分类。

内部效度。本研究采用的来源选择方法详述于第 4.1 节。为确保研究可复现，我们仔细识别并报告了用于检索同行评审文献的书目来源、用于检索灰色文献的搜索引擎、检索式，以及纳入和排除标准。选择过程中可能出现的问题，与搜索词的选择有关，这些搜索词可能导致结果集不完整。为降低这一风险，我们扩展了 Recupito 等人采用的检索式，加入了 MLOps 一词的可能同义词。为克服搜索引擎的局限性，我们从六个书目来源检索了学术文献。来源；同时，我们纳入了来自 Google、Medium Search、Twitter Search 和 Reddit Search 的灰色文献。此外，我们采用了滚雪球过程，以纳入所有可能的来源。纳入与排除标准的应用可能会受到研究人员意见和经验的影响。为减轻这一威胁，所有来源均由至少两位作者独立评估。此外，为评估评审者间一致性的质量，我们计算了 Cohen's Kappa 系数，结果显示一致性几乎达到完美。我们认为，如此高的一致性源于纳入与排除标准易于应用；尤其是，排除不包含任何工具相关内容的论文十分容易。

结论效度。结论效度与根据结果得出的结论的可靠性有关 [9]。为确保处理过程的可靠性，我们没有自行定义术语，而是依据现有的开发运维和机器学习运维步骤进行分类。此外，所有主要来源至少由两位作者审阅，以减轻数据提取中的偏差；每项分歧均通过协商达成共识，必要时由第三位作者参与。

外部效度。 外部效度与多声部文献综述结果的可推广性有关。在本研究中，我们绘制了有关机器学习运维工具的文献图谱，兼顾学术文献和灰色文献。然而，我们无法声称已筛查所有可能的文献，因为某些文档可能未被妥善编入索引，也可能受版权保护，甚至无法免费获取。

8. 相关工作

多项研究强调了当前辅助机器学习流水线的工具现状。Idowu 等人 [13] 根据特性分析，对 30 种机器学习资产管理解决方案进行了比较。研究结果表明，支持实验可复现性所需的跟踪、探索和检索活动的机器学习系统，已得到实践现状和研究现状的支持。Ruf 等人 [14]通过评估可能支持机器学习流水线的 26 种机器学习运维工具，并考察其潜在支持能力，展示了机器学习运维的潜力。研究结果显示，没有任何机器学习运维解决方案能够实现足够程度的自动化来支持机器学习运维工作流，并且若干产品的功能存在重叠，提供了相同的支持。Kolltveit 等人 [15] 开展了一项系统性文献综述，以调查实际用于辅助模型运维化的工具，旨在识别功能缺口。综述结果表明，管理机器学习模型的边缘部署，以及实现动态模型切换和持续模型监控等功能，仍是机器学习运维面临的真正挑战。Testi 等人[16]开展了一项综述，旨在建立机器学习运维分类体系，概述用于构建机器学习流水线的方法和流程。作者总结了根据其所发现的工具集合，在机器学习流水线的特定阶段采用某一选定工具的优势和劣势，

最后，Recupito 等人 [1] 对当前研究和实践中的多种机器学习运维工具及其功能进行了多声部文献综述。在提取出 13 种机器学习运维工具后，对涉及整个机器学习流水线、涉及模型管理以及涉及数据管理的各个方面进行了特性分析和比较。

本研究考察的大多数最新综述都强调了弥合研究与实践之间差距的必要性。本研究的结果纳入了从业者用于自动化和维护机器学习系统的 84 种工具。此外，我们通过考察这些工具之间的兼容性，进一步研究了机器学习运维工具的现状，使从业者能够选择理想的解决方案组合，以便在机器学习系统中采用机器学习运维。

9. 结论

在本研究中，我们开展了一项多元文献综述，对开发运维流程中的机器学习运维工具进行分类，并识别工具之间可能存在的不兼容情况。

我们从 254 项主要研究中提取了 84 种工具（其中 203 项主要研究来自灰色文献，51 项主要研究来自白色文献）。最后，我们提出了开发运维流程中机器学习运维工具的图形化表示，这有助于研究人员和从业者快速了解现有的机器学习运维工具。值得注意的是，绝大多数端到端机器学习运维平台通常无法集成其他工具，而其他工具之间则未发现不兼容情况。

因此，可以将不同开发运维阶段的所有工具整合到一条无缝流水线中。

我们计划通过开展一项工业调查来扩展这项工作，以探究业界采用的最常见流水线，以及每种工具所带来的感知收益、问题和实用性。

致谢

本研究部分由 IndustryX 项目（Business Finland）、芬兰科学院的 MuFAno 项目（资助编号 349488）、EME-LIOT 项目（意大利，MUR - PRIN 2020 计划，合同编号 2020W3A5FY）以及瑞士国家科学基金会通过 SNF 项目（编号 PZ00P2 186090）资助。

CRediT 作者贡献声明

Sergio Moreschini： 概念化、调查、数据整理、撰写——初稿、撰写——审阅与编辑、可视化。

Gilberto Recupito： 调查、数据整理、撰写——审阅与编辑。

Valentina Lenarduzzi： 撰写——原始草稿、撰写——审阅与编辑、可视化。

Fabio Palomba： 监督、审阅、编辑。

David H¨astbacka： 概念化、方法论、监督、审阅、编辑、经费获取。

Davide Taibi： 概念化、方法论、监督、审阅、编辑、经费获取。

参考文献

[1] G. Recupito、F. Pecorelli、G. Catolino、S. Moreschini、D. Di Nucci、F. Palomba、D. A. Tamburri，《机器学习运维工具及其特性的多声部文献综述》，欧洲微电子软件工程与高级应用会议（2022）。[2] J. Carver、N. Juristo、M. T. Baldassarre、S. Vegas，《软件工程实验的复制研究》，《实证软件工程》（2014）267–276。[3] L. Bass、I. Weber、L. Zhu，《开发运维：软件架构师的视角》，Addison-Wesley Professional，2015。[4] S. Moreschini、F. Lomio、D. H¨astbacka，《面向可演进人工智能密集型软件系统的机器学习运维》，载于：2022年IEEE国际软件分析、演化与再工程会议（SANER），第1293–1294页。[5] V. Garousi、M. Felderer、M. V. M¨antyl¨a，《在软件工程中纳入灰色文献并开展多声部文献综述的指南》，《信息与软件技术》106（2019）101–121。[6] C. Wohlin，《系统文献研究中滚雪球检索的指南及其在软件工程中的复制研究》，载于：软件工程评估与评价国际会议，Ease '14，第1–10页。[7] B. Kitchenham、S. Charters，《软件工程系统文献综述实施指南》，技术报告 Ebse 2007-001，基尔大学，2007。[8] B. Kitchenham、P. Brereton，《软件工程中系统综述过程研究的系统综述》，《信息与软件技术》55（2013）2049–2075。[9] C. Wohlin、P. Runeson、M. H¨ost、M. C. Ohlsson、B. Regnell，《软件工程实验》，2012。[10] G. A. Center，《机器学习运维：机器学习中的持续交付与自动化流水线》，https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning#devops_versus_mlops 2022，.

[11] Docker，《什么是容器？》，https://www.docker.com/resources/what-container/ 2022，.

[12] A. Felix，《如何使用 Prometheus 和 Grafana 监控 Jenkins 指标？》，https://medium.com/@AnnFelix/how-to-monitor-jenkins-metrics-using-prometheus-grafana-152a98d6c7a6，2022。[13] S. Idowu、D. Str¨uber、T. Berger，《机器学习中的资产管理：研究现状与实践现状》，《ACM计算调查》55（2022）。[14] P. Ruf、M. Madan、C. Reich、D. Ould-Abdeslam，《揭开机器学习运维的神秘面纱并提出开源工具选择方案》，《应用科学》11（2021）。[15] A. B. Kolltveit、J. Li，《机器学习模型的运维化：系统文献综述》，载于：2022年IEEE/ACM第一届负责任人工智能软件工程国际研讨会（SE4RAI），第1–8页。[16] M. Testi、M. Ballabio、E. Frontoni、G. Iannello、S. Moccia、P. Soda、G. Vessio，《机器学习运维：分类体系与方法论》，《IEEE Access》10（2022）63606–63618。