“我们不应该做的侦探工作”：机器学习系统中符合监管要求的数据质量方面从业者面临的挑战

Yichun Wang阿
姆斯特丹大学 荷兰

Kristina Irion 阿
姆斯特丹大学 荷兰

Paul Groth 阿姆
斯特丹大学 荷兰

Hazar Harmouch
阿姆斯特丹大学 荷兰

arXiv:2602.05944v1 [cs.DB] 5 Feb 2026

摘要

随着监管要求不断扩展，确保机器学习（ML）系统中的数据质量变得日益复杂。在欧盟，诸如《通用数据保护条例》（GDPR）和《人工智能法案》（AI Act）等框架提出了与机器学习实践中的技术关注点密切对应的数据质量要求，同时也将要求延伸至与问责制、风险管理和人权保护相关的法律义务。本文介绍了一项定性访谈研究，研究对象是在受监管场景中从事机器学习系统工作的欧盟数据从业者。通过半结构化访谈，我们探究从业者如何理解符合监管要求的数据质量、他们面临哪些挑战，以及他们认为哪些支持必不可少。研究结果揭示，法律原则与工程工作流之间存在持续差距，数据管道彼此割裂，现有工具能力有限，技术团队与法律团队之间的责任边界不清，并且质量实践往往倾向于被动应对、由审计驱动。我们还发现，从业者需要具备合规意识的工具、更清晰的治理结构，以及推动数据治理转向主动管理的文化转变。

CCS 概念

• **社会与专业主题** → **政府法规**；**质量保证**； • **以人为本的计算** → **用户研究**； • **计算方法** → 机器学习。

关键词

数据质量、机器学习、监管要求、合规、《人工智能法案》、《通用数据保护条例》、数据从业者、欧盟

特此免费授予许可，允许为个人或课堂使用而复制本作品的全部或部分内容（无论数字版还是纸质版），但不得以营利或商业利益为目的制作或分发副本，且副本必须保留本声明及首页上的完整引文。作者以外的其他方拥有的本作品组件版权必须予以尊重。注明出处后可以摘录。若要以其他方式复制、再版、发布到服务器或分发到列表，必须事先获得具体许可和/或支付费用。请通过 permissions@acm.org 申请许可。*Conference'17*，美国华盛顿特区 ©2026 版权所有者/作者保留版权。出版权授权予 ACM。ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MMhttps://doi.org/10.1145/nnnnnnn.nnnnnnn

ACM 引用格式：
Yichun Wang、Kristina Irion、Paul Groth 和 Hazar Harmouch。2026 年。《“我们本不该做的侦探工作”：机器学习系统中符合监管要求的数据质量所带来的从业者面临的挑战》。收录于 。 ACM，美国纽约州纽约市，14 页。
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 引言

数据质量对于机器学习系统的可靠性、公平性和安全性至关重要 [24, 33, 60]。低质量数据可能导致有偏差的结果、不稳定的模型以及影响个人和组织的系统故障 [48, 54, 65]。在许多技术研究中，机器学习中的数据质量是围绕性能和鲁棒性来界定的。从业者致力于避免标签噪声、分布偏移、数据漂移以及会降低模型准确性或稳定性的不一致问题 [8, 17, 50, 52]。为此，他们会部署验证检查、监控和流水线编排工具，以确保数据管道在生产环境中保持稳定 [64, 75]。

除了这些技术层面的考量，在欧盟，数据质量也是一项监管要求，与问责制、风险管理以及基本权利保护息息相关。欧盟的《通用数据保护条例》（GDPR）和《人工智能法案》（AI Act）等法规对数据质量提出了明确要求 [20, 21]。《通用数据保护条例》确立了准确性、数据最小化、目的限制和问责制等原则，适用于个人数据的处理 [28, 34]。《人工智能法案》进一步要求高风险人工智能系统采用适当且具有代表性的训练、验证和测试数据，并采取偏差缓解措施、提供文档以及持续进行风险管理 [42, 49, 69]。总而言之，这些法规要求数据质量实践同时满足技术和法律层面的要求。

在实践中，使数据质量符合监管要求并非易事。现代机器学习系统依赖复杂的数据管道，涵盖数据摄取、数据转换、特征工程、训练、部署和监控等环节 [8, 53, 58, 65]。数据质量责任分散在不同团队、工具和组织边界之间 [35, 46, 51, 60]。法律要求通常以高层次原则的形式表述，而工程师面对的则是具体约束、指标和交付时间表 [5, 18, 51]。这种结构性错位，使得将监管要求转化为贯穿多个数据管道组件及组织环境的具体数据质量实践变得更加复杂。为明确这一挑战，我们将“符合监管要求的数据质量”概念化为：技术数据质量实践在满足工程

标准并履行适用的法律义务方面所达到的程度。我们系统地将常用的数据质量维度与《通用数据保护条例》和《人工智能法案》中的相应要求进行映射（见附录 A 表 2）。然而，从业者如何在实践中落实这种对齐，仍缺乏深入研究。因此，从业者常常不确定应如何在日常数据质量保障实践中解读并落实监管要求 [40, 56, 66, 71]。关于他们如何应对这一领域，尤其是在欧盟背景下如何进行应对，目前仍缺乏实证性认识。

本文通过从从业者的视角审视符合监管要求的数据质量来填补这一空白。从业者在日常数据工作中落实监管要求方面发挥着核心作用，但他们对符合监管要求的数据质量的看法仍未得到充分探讨。我们提出以下研究问题：

从业者在确保机器学习系统中符合监管要求的数据质量方面面临哪些现实挑战，以及他们在实践中需要哪些形式的支持来应对这些挑战？

为有效回应这一核心问题，本研究将深入探讨以下具体子问题：

- **研究子问题1：** 从业者目前如何在日常机器学习工作流中理解并落实关键的法规适配的数据质量维度，以及哪些因素塑造了这些实践？
- **研究子问题2：** 如今，从业者使用哪些工具、方法和基础设施来管理符合监管要求的数据质量？他们认为还需要哪些额外能力才能满足监管义务？
- **研究子问题3：** 技术团队与法律或合规团队之间的协作模式，如何影响符合监管要求的数据质量的实施？

为回答这些问题，我们对14位欧盟从业者进行了半结构化访谈，他们从事处理个人数据或在受监管领域运行的机器学习系统相关工作。参与者涵盖多个职位，包括数据工程、数据科学、机器学习工程和合规。访谈围绕九个数据质量维度展开，包括准确性、可追溯性、相关性和无偏差（见图 1）。我们请参与者回顾他们在应对法律合规要求所引发的数据质量问题方面的经验。此外，我们还介绍了三个受《通用数据保护条例》和《人工智能法案》要求启发而设计的基于场景的短案例（见附录 C）。这一设计鼓励参与者结合具体实践进行反思，而非停留在抽象原则层面。

本文主要有三项贡献。首先，本文实证考察了欧盟从业者如何解读并在机器学习流水线中落实符合监管要求的数据质量。其次，本文指出了影响这项工作的跨领域挑战，包括将法律原则转化为工程任务、数据管道碎片化、现有工具的局限，以及技术团队与法律团队在协作和责任分担方面的缺口。第三，本文提炼了从业者提出的合规感知数据质量未来研究方向，包括探索新型技术基础设施、治理结构和操作指南的可能性。图 1概述了研究设计和分析结果。

2 文献综述

本研究处于数据质量与欧盟监管的交汇处。以往研究提出了丰富的技术方法、治理框架和法律分析，但这些研究方向往往各自独立发展。本节从技术、监管和实证视角回顾主要文献，并指出其中的空白，以此说明开展本项以从业者为中心的研究、探讨实践中的符合监管要求的数据质量的动机。

2.1 数据质量与机器学习：技术视角

传统上，数据质量在数据管理领域通过准确性、完整性、及时性、一致性、相关性和可访问性等维度进行研究 [6, 73]。在机器学习中，数据质量日益被视为决定模型性能和鲁棒性的首要因素。近期研究强调以数据为中心的人工智能，指出系统性地改进数据，例如进行整理、扩大覆盖范围和确保标签完整性，往往比提高模型复杂度带来更大的收益 [60, 76]。研究表明，标签噪声 [48], 、数据集偏移 [54], 、类别不平衡 [37], 和代表性 [11] 都会影响模型行为和泛化能力。

机器学习流水线中的质量保证实践体现了这些关注点。团队通常采用模式验证、基于规则的检查和统计检验来检测异常和漂移 [8, 64]。特征存储和数据版本控制系统旨在通过控制数据依赖关系，稳定训练和推理过程 [22]。实验跟踪框架会记录配置和指标，以支持可复现性 [75]。Great Expectations、Deequ 等工具和类似框架让团队能够在代码中声明并测试数据质量约束 [29, 63]。这些实践如今已成为机器学习运维中的标准做法。然而，它们主要着眼于提升预测性能或系统可靠性，并不总能满足法律合规相关的要求。

除了质量检查之外，技术治理机制也应运而生。数据目录、数据沿袭系统和溯源框架会追踪数据在各条数据管道中的流动与转换 [12, 32]。元数据管理用于描述数据集、特征和模型 [62]。数据集数据表和模型卡等文档工件旨在提高透明度和可审查性 [24, 47]。Apache Atlas、Amundsen、DataHub 以及商业数据目录等系统旨在提供数据资产的端到端可视性 [2, 30, 41]。MLflow、Kubeflow 等机器学习运维平台以及云原生解决方案支持实验跟踪、模型版本管理和部署 [38, 74]。这些工具和实践共同构成了用于质量保证与治理的技术栈。

然而，重要的缺口依然存在。数据质量被概念化为可在技术栈中加以检测和监测的属性。它刻画数据如何流动以及模型如何接受训练。它很少将目的限制、合法依据、同意或保留约束等法律要求编码进去。跨派生特征、缓存和历史训练集的删除传播也很薄弱 [13, 26, 36]。公平性与子群体评估往往依赖难以复现的临时脚本 [33, 46, 55]。文档常常是手工编写的，并会偏离持续演进的流水线，在快速变化的系统中尤其如此 [27, 31]。工具之间的碎片化会在端到端监督中造成盲区。因此，技术质量与监管质量之间的差距被内置于许多当前

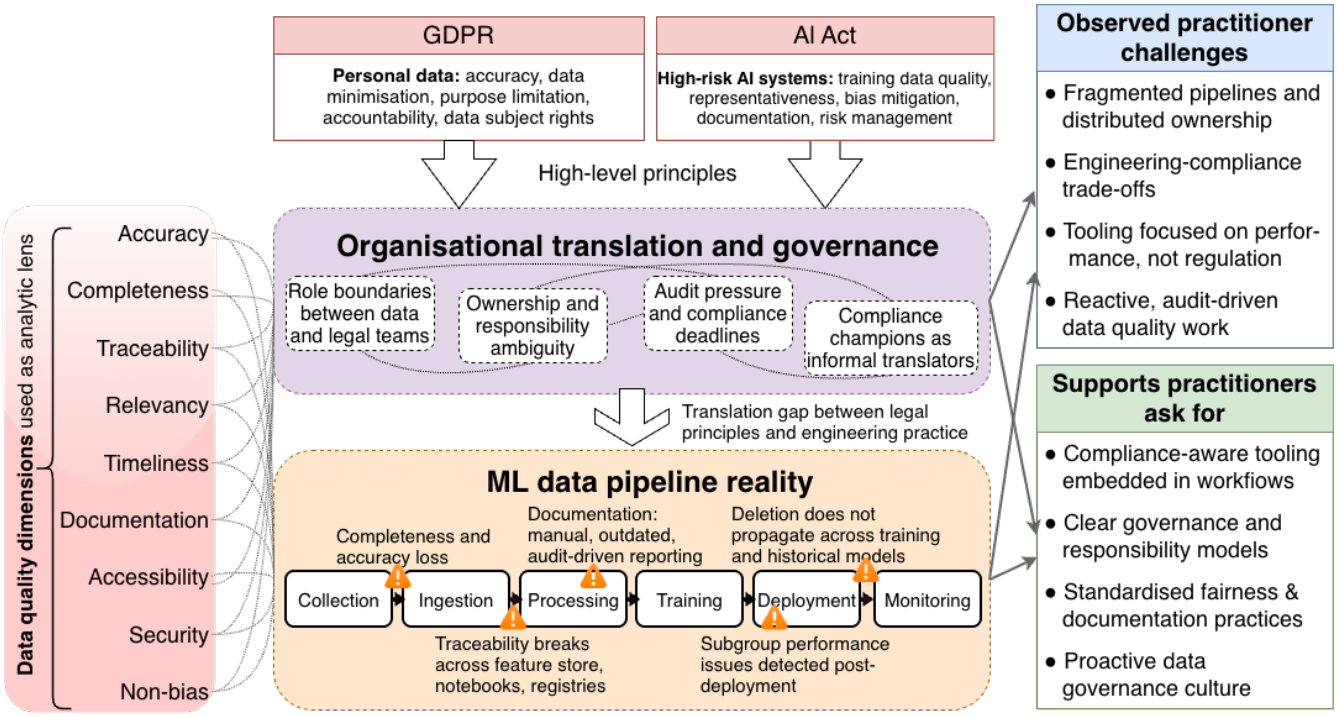


图1：实践中的符合监管要求的数据质量——挑战与期望特性

方法之中。我们的研究结果印证了这一系列工作，并提供了实证证据，说明此类缺口如何在从业者的日常合规工作中显现。

2.2 关于数据质量的监管视角

欧盟法规对数据质量提出了明确要求，尽管这些要求处于较高的抽象层次。《通用数据保护条例》通过准确性、数据最小化、目的限制、存储限制、完整性与保密性以及问责制等原则，对包含个人数据的数据集处理进行监管 [20]。这些原则直接影响个人数据的收集、处理、保留和文档记录方式。准确性关系到记录正确性和更正义务。数据最小化和目的限制约束特征选择与数据再利用。存储限制决定保留与删除。完整性与保密性推动访问控制和保护措施。问责制要求提供能够将各项选择与法律义务联系起来，并证明遵守这些义务的证据和文档。数据主体权利进一步提出了对可追溯性、更正、删除权和可携带性的技术要求 [68]。这些权利将技术血缘转化为合规证据。它们还暴露出那些未被设计为支持此类义务的数据管道中的薄弱环节。

《人工智能法案》将这些要求扩展至高风险人工智能系统。它强制要求数据治理和管理实践、有代表性且已缓解偏见的训练数据、风险管理、文档记录以及上市后监测 [21]。法学研究已分析这些条款的规范目标及其对人工智能可问责性的影响 [43, 69]。然而，这些义务是基于原则且技术中立的 [23]。它规定了目标，但将实施选择保持开放。

将法律要求应用于具体的机器学习项目，需要根据具体案例进行解读，这造成了持续存在的转译问题 [7, 61, 69, 71]。法规阐明原则，而从业者需要具体的行动和工件。以往的跨学科研究凸显了在技术系统中落实法律规范的难度 [5, 18, 19, 25, 70]。研究指出，工程师往往缺乏充分指导，只能自行解读合规义务 [3, 66, 72]。正如我们稍后展示的研究发现所示，机器学习从业者明确表示，他们需要知道哪些流水线步骤、工件和检查能够提供充分证据。他们需要法律团队与技术团队之间的共同语言。

2.3 从业者经验与组织动态的实证研究

越来越多的实证研究开始考察机器学习从业者如何践行负责任人工智能实践。访谈和调查研究记录了与公平性、可解释性和问责制相关的挑战 [33, 45, 57]。这些研究表明，责任往往分散在各个团队之间，激励机制彼此不一致，而伦理实践也常常到开发周期后期才得以应用。

关于治理和溯源采用的研究揭示了类似的模式。从业者重视可追溯性，但难以在组织边界和异构工具链之间维持这种可追溯性 [59, 60]。问责制和规范性考量很少融入日常开发工作流程，而往往是在事后才浮现

等关键技术决策已经做出之后 [51]。实际上，治理责任往往落在非正式合规倡导者身上，由他们在法律期望与技术工作流程之间进行协调，但这些职位通常资源不足，也缺乏正式职权 [46]。

然而，现有的实证研究往往聚焦于公平性或可解释性等模型层面问题 [67]。数据质量很少被作为贯穿数据摄取、数据转换、训练、部署和监控的跨领域问题加以考察。欧盟特定监管环境也未得到充分体现。因此，对于从业者如何应对具体的《通用数据保护条例》和《人工智能法案》要求，将数据质量的各个维度付诸实践，人们知之甚少。本研究直接回应了这一空白。

2.4 研究空白

技术文献为验证、监控、版本管理和数据沿袭提供了有力的方法，但其主要取向仍是性能和运营稳定性。监管及社会法律研究从较高层面阐述了数据保护原则，却很少说明如何在具体的机器学习项目中将这些要求付诸实践。实证研究揭示了组织与文化障碍。总体而言，这些研究方向涉及相关问题，但彼此之间的联系仍然薄弱。

本研究通过考察在欧盟工作的从业者如何在实践中应对这一领域，为相关研究作出了贡献。这些从业者处理个人数据或高风险机器学习系统。通过定性访谈，我们探究他们如何理解关键的数据质量维度、在使工作符合监管要求时遇到哪些挑战，以及他们认为还缺少哪些支持。这一视角补充了技术和法律层面的研究方法，并为设计未来的工具、流程和政策提供了实证依据。

3 方法论

我们采用定性访谈研究设计。这种方法广泛用于计算机科学研究，以探究从业者在现实世界机器学习开发情境中的经验与权衡 [33, 39, 44]。我们采访了十四位在欧盟工作的从业者。在本研究中，我们对从业者作了广义界定，指参与开发基于机器学习的产品或服务任何阶段的人员。样本涵盖运营职位（如数据收集员和数据标注人员）、技术职位（如数据工程师、数据科学家和机器学习工程师），以及法律或合规职位（合规专员和法律工作人员）。表 1 概述了参与者的特征。

3.1 访谈设计

半结构化访谈协议包含四个主要部分（见附录 B）。首先，我们询问参与者的工作背景和项目设置，包括他们日常工作中现有的数据质量标准。其次，我们讨论了一个因合规风险而显现的数据质量问题的具体案例。第三，我们提供了一组简短的情境短文 [15],，以了解参与者在特定监管或数据质量情境下会如何应对。最后，对于具有丰富合规经验的受访者，我们还询问了他们与法律团队协作的相关问题，他们感知到的挑战、权衡和动机。访谈方案根据每位参与者的职位和专业知识进行了调整。

设计与合规相关主题的访谈需要谨慎。受访者可能会因担心社会期许偏差或感知到组织风险而不愿畅所欲言。为减轻这种压力，我们采取了几项策略。我们明确说明，所有涉及个人、组织、产品和服务的可识别信息都会匿名化。在安排正式访谈之前，我们与大多数参与者进行了非正式交流，以建立信任并减轻他们的顾虑。访谈过程中，我们使用中性且专业的语言，以尽量减少参与者的不适。我们强调，本研究旨在了解挑战和支持需求，而非评估合规表现。我们避免使用任何可能暗示责备或不当行为的措辞。我们提醒参与者，他们可以随时暂停或停止访谈或录音。这种做法有助于营造安全的环境，鼓励参与者坦诚而深入地作答。完整的访谈指南见附录 B。

作为这项审慎设计的一部分，我们采用情境短文作为一种抽象的、基于情境的引导式 elicitation 技术。情境短文描述了与参与者所在组织无关的假设情境，有助于减少防备心理，同时促使参与者反思现实中的实践与权衡。既有研究表明，在伦理或法律敏感领域，情境短文有助于引出隐性推理和可行性约束 [4, 16]。这些情境短文旨在反映《通用数据保护条例》和《人工智能法案》中反复出现的监管冲突，包括偏差和子群体表现（《通用数据保护条例》关于个人数据特殊类别的规定，第22条第4款；《人工智能法案》关于数据治理的规定，第10条第2、3和5款）、数据集来源不明确（《通用数据保护条例》关于合法依据的规定，第6条；《人工智能法案》关于数据治理的规定，第10条），以及审计准备（《通用数据保护条例》关于问责制的规定，第5条第2款；《人工智能法案》关于文档的规定，第18条）。我们明确鼓励参与者描述其团队在现实中会如何应对，包括哪些做法不可行或需要额外支持。完整的情境短文见附录 C。

3.2 招募与实施

我们通过便利抽样招募参与者，主要经由专业网络以及线下的机器学习和人工智能活动，包括会议、聚会和研讨会。为确保对《通用数据保护条例》和《人工智能法案》均具有相关性，我们有意纳入工作涉及个人数据（如《通用数据保护条例》第4条第(1)款所定义 [20]）或高风险机器学习系统（如《人工智能法案》附件III所列[21]）的参与者。招募与访谈于2025年4月至12月期间迭代进行。数据收集与分析同步推进，使新出现的主题能够为持续招募提供依据。当反复出现的主题在分析中趋于稳定、表明已达到主题饱和时，我们停止招募。

访谈根据参与者的偏好，以面对面或视频会议的方式进行。每场访谈约持续一小时。我们录音以便转录。转录过程中，我们删除了所有个人及组织标识信息，并在之后删除了原始音频文件。所有访谈记录和笔记均安全地存储在本地加密设备上。

表1：访谈受访者概览，包括行业领域、组织规模、职位和核心职责。

	行业	公司规模	Role	主要职责
R1	金融（银行业）	>10,000	数据科学家 / 数据工程师-neer	用于监控的机器学习模型开发
R2	金融（银行业）	>10,000	数据科学家	合规 用于合规的机器学习模型验证
R3	金融、电信-通信	>10,000	数据工程师	数据工程（自由职业）
R4	零售业	>10,000	机器学习工程师	销售预测模型
R5	金融、交通运输	>10,000	数据科学家	欺诈预测建模
R6	交通运输	500–1,000	数据分析师	人力资源分析仪表板
R7	医疗健康	1,000–5,000	数据科学家 / 研究人员	疾病预测模型
R8	市场研究	25–100	数据科学家	使用大语言模型分析访谈数据
R9	生成式人工智能、交通运输	>10,000	数据科学经理	数据科学监督与合规
R10	金融、地理学	1,000–5,000	数据科学家 / 数据收集员	房价预测
R11	审计	1,000–5,000	法律与合规	合规职能
R12	审计	1,000–5,000	法律与合规	合规职能
R13	审计	1,000–5,000	法律与合规	合规职能
R14	审计	1,000–5,000	法律与合规	合规职能

3.3 分析

我们遵循 Deterding 和 Waters 所述的迭代编码方法 [14], , 结合归纳式和演绎编码策略。我们首先根据既有的数据质量维度、相关欧盟法规（《通用数据保护条例》和《人工智能法案》）, 以及研究问题中的主题，制定演绎编码。我们将这些编码系统地应用于访谈记录和田野笔记。同时，我们也对初始框架未预料到的新兴主题保持开放态度。随后，我们进行了第二轮分析性编码，以识别不同参与者之间的模式、相似之处和差异。通过与机器学习、人工智能和法律领域的研究人员讨论，我们进一步完善了新兴主题。这一迭代过程使我们得以识别意义模式，并以此为最终结果提供依据。

3.4 局限性

我们在整个研究过程中都谨慎地尽量减少偏差，包括访谈设计、招募和分析阶段。不过，仍需承认若干局限性。

首先，招募过程带来了抽样限制。我们力求涵盖不同领域、职位、国家和组织规模的参与者，但参与者总数仍然有限。样本中的男性和大型企业从业者比例也偏高。这反映了人工智能从业者队伍的整体趋势：男性比例偏高，且从业者集中于大型企业 [1]。若能在更长时间内收集更广泛的样本，将有助于提升代表性。尽管如此，该样本量符合定性研究的惯例，例如，ACM CHI 研究的平均样本量为 12 [10]。此外，机器学习和数据工程领域发展迅速。参与者对新兴技术或工具的理解可能会过时。尽管如此，他们描述的许多根本需求，例如更清晰的治理以及对监管对齐的支持，可能会持续存在。

其次，本研究采用访谈方法，捕捉的是对组织行为的看法，而非直接观察。参与者的陈述可能将事实描述与假设或不完整信息混杂在一起，而这些内容会受到其经验、知识水平和组织背景的影响。因此，可能存在自我报告偏差和感知偏差。我们也无法直接访问内部系统或组织工作流程，这限制了我们核实其陈述或结合具体情境进行解读的能力。田野研究或民族志观察可以补充这一方法，并为符合监管要求的数据质量在实践中如何落实提供更丰富的证据。

总体而言，本研究应被视为理解与监管要求协作的数据从业者所面临挑战与需求的初步一步。它呈现了当前实践的概况，并凸显了未来研究的方向。

4 访谈结果

研究中的参与者普遍对监管要求有所了解，尤其是与《通用数据保护条例》相关的要求。这并不令人意外，因为受访者均因至少具备一定的合规要求应对经验而被选中。许多人表示，《通用数据保护条例》已成为其工作中的常规部分，而他们对《人工智能法案》的了解则远不够一致。不同参与者所属的行业也存在知识差异。来自银行业或金融业的参与者表示，他们对监管框架的熟悉程度高于来自监管较少行业的参与者。从业者描述了推动其数据质量和合规实践的一系列动机。大多数参与者受到审计、法律义务以及担心罚款等外部压力的驱动。参与者 R6 表示：“《通用数据保护条例》的罚款是真实存在的。作为一家小公司，如果我们被处以巨额罚款，很可能会破产，所以我们格外谨慎。”其他人则受到内部价值观的驱动，包括组织文化、领导层期望和个人责任。正如受访者 R5 所说，

> “作为客户，我不希望自己的数据被草率对待。因此，我在处理用户的个人数据时会更加谨慎。”

这些动机与若干影响优先级的因素相互作用，包括所处理数据的敏感性、组织在数据治理方面的成熟度、工具的可用性，以及该行业是否受到严格监管。

尽管人们普遍意识到合规的重要性，参与者仍表示，他们对于如何将法律原则转化为日常工程任务，以及如何应对由此产生的权衡，存在很大不确定性。以下主题介绍了他们报告的现实挑战，以及他们认为必不可少的支持。

4.1 确保数据质量符合监管要求的现实挑战

数据管道碎片化及技术难题阻碍了数据质量实践的一致性。参与者经常强调，他们面临的许多符合监管要求的数据质量挑战，直接源于现代机器学习流水线中数据流转方式的碎片化。许多人表示，当数据跨越组织边界流转时，负责上游数据摄取和工程工作的团队往往缺乏下游建模团队所掌握的业务背景或领域知识。这种脱节常常损害数据的准确性和完整性。正如从业者 R7 所解释的，上游团队 " 并不总是知道模型真正看重什么 "，这使他们难以设定恰当的验证约束，也难以识别重要属性是否缺失或有误。另一位参与者 R9 总结道："数据从上游团队开始流转，等到数据到我们手上时，已经没人真正知道谁验证过什么了。"当基础质量检查彼此不一致时，下游团队便会接手无法完全信任的数据，却仍须负责将其用于受监管的机器学习系统。

这种碎片化也让可追溯性 变得脆弱。几位从业者表示，数据进入已经完成的机器学习流水线后，他们便无法再了解此前的数据转换过程。这使得数据经过摄取服务、数据转换流水线、特征存储和已部署模型时，可追溯性"充其量也只是部分具备"。即使存在局部的日志记录实践，也很少能串联起从原始来源到模型输入的完整路径，因此难以证明符合可审计性或溯源方面的要求。一位受访者 R2 开玩笑说，如果他们必须重建数据集从源系统到最终模型的完整生命周期，就得"追着五个不同的团队和七种不同的工具跑"。文档 实践也反映了这种碎片化。参与者表示，Git、笔记本和内部维基中的文档"各说各话"，而且没有机制能在数据沿流水线更新时让这些来源保持一致。从业者指出，这些缺口在审计审查期间尤为明显，正如参与者 R10 所说，团队会"意识到每个人的文档说的都略有不同"。碎片化也影响可访问性 和安全性，因为数据权限通常由基础设施团队而非机器学习团队管理，导致某些领域的访问权限过于宽泛，而其他领域则出现瓶颈。

当监管要求删除个人数据时，这些挑战尤为突出。参与者反复指出，删除请求无法可靠地沿机器学习流水线传递。即使可以从主表中删除一条记录，参与者解释说，这些数据的痕迹仍会贯穿机器学习生命周期，留存在派生特征、历史数据集和中间提取数据中。正如一位工程师 R5 所说，

> "如果用户要求删除，我们该如何将其移除
> 从派生特征或历史模型训练
> 数据中？重新训练的成本会变得高昂，甚至无法实施。"

在各个组织中，即使是积极主动的团队，也很难在没有任何人对数据的整个生命周期负责时，维护数据质量和合规要求。结果是，虽然局部存在良好实践，却始终无法形成监管制度所设想的连贯、端到端治理 [60]。

工程目标与监管数据质量要求之间的权衡。参与者描述了模型性能表现与模型合规性之间经常存在的紧张关系。通过情境短文，从业者被引导去思考，在实际环境中，技术决策如何与公平性、问责制以及文档要求发生冲突。这些情境帮助揭示了从业者如何在现实约束下应对这些权衡。

许多参与者表示，在部署后处理**偏差和子群体表现**是最困难的挑战之一。几位参与者解释说，模型在总体指标上往往表现良好，却掩盖了特定群体表现显著下降的问题。他们指出，弥补这些差距需要的不只是重新训练。团队还需要进行额外的数据收集、重新设计特征，或采用专门的评估流程。然而，在严格的交付时间表下，这些步骤很难获得合理 justification。一位工程师 R1 说道，

> "我们想修复它，但业务方仍希望模型保持上线。
> 这变成了一场谈判，而不是一项质量决策。"

参与者还指出，公平性修正可能会降低模型对总体人群的准确性。一些人指出，公平性修正很少有明确的验收标准。这迫使团队在性能损失与满足《人工智能法案》对监控和无偏差的要求之间进行权衡。这种权衡很少是直截了当的。

关于**数据集来源不明确**的情景片段引发了强烈反应。从业者表示，他们经常遇到前景看好但文档不完整的第三方数据集。许多人想使用这些数据，因为它有望带来性能提升，但他们也认识到其中的监管风险。几位受访者表示，他们会抵制管理层的压力。他们会阻止使用该数据集，直到合法依据和收集做法得到澄清。另一些人承认，在快节奏的环境中，交付压力有时会占上风。R4表示，他们可能会用该数据集开展实验，"只是想看看它是否有帮助"，尽管他们知道这会造成合规债务，因为这不符合《通用数据保护条例》对合法依据的标准，也不符合《人工智能法案》治理。工程师R3说："该数据集或许能帮助模型，但如果我们无法信任其来源，它就会成为一颗定时炸弹。"这一情景揭示了一种更深层的模式。性能提升备受重视，但来源核查缓慢且不确定。团队夹在创新激励与问责制要求之间。

关于**审计准备**的情景案例进一步凸显了权衡取舍。参与者表示，他们可以提供一些证据，例如准确性检查、基本的数据沿袭或偏差测试示例。然而，许多人坦言，他们难以出具《通用数据保护条例》问责制或《人工智能法案》所要求的完整文档

文档保存。他们缺乏一条连贯的端到端叙述，将数据收集与模型部署联系起来。参与者R3表示，他们只能展示“我们上个月测试了什么，而不是流水线每一步发生了什么”。该情景案例揭示了一个关键洞见。团队知道如何监测、验证和衡量性能，但缺乏将这些活动转化为可供审计的合规证据的系统。

从这些情境短文中，显现出一个清晰的模式。从业者理解法规要求，但他们所处的环境中，工程目标、时间压力和组织结构共同决定了哪些事情切实可行。这些权衡并非疏忽的表现，而是反映了现实：合规与工程受到不同激励因素和时间安排的驱动。参与者描述了持续不断的平衡过程：他们既要满足法律标准，又要确保系统保持可用并具备竞争力。

缺乏支持符合监管要求的数据质量的工具。参与者指出，日常数据质量管理所用工具与监管框架提出的要求之间，差距日益扩大。他们的团队依赖商业平台、开源库和定制系统，但这些工具大多面向功能性或性能导向的质量，很少能支持欧盟法规所要求的问责制、可解释性或可审计性。因此，从业者必须通过人工工作弥补不足，尤其是在准备合规文档或风险评估时。

参与者解释说，商业工具往往过于僵化，无法体现领域特定约束，而开源解决方案虽能提供有用的组件，但面向合规的功能却很少。许多参与者表示，他们已构建内部验证器或监控层来加以弥补。正如工程师 R3 所说，他们的定制工具成了“我们理解自身数据所需的黏合剂，因为没有任何现成工具理解我们的领域，或合规对我们意味着什么。”其他人解释说，他们创建内部系统只是为了跟上产品开发的节奏，并指出只有大型组织才有资源这样做。对于规模较小的公司，跟上合规要求被描述为更具挑战性。

工具方面的不足延伸到了更广泛的机器学习基础设施。MLflow、Azure ML 和 SageMaker 等机器学习运维平台因版本管理和可复现性而受到重视，但参与者批评它们缺乏符合监管要求的数据质量所需的功能，例如记录处理目的、法律依据或保留规则。版本管理工具能够记录什么发生了变化，却不能记录为什么，而数据沿袭系统展示的是技术转换，而非其合规影响。几位参与者指出，模型版本之间的细微变化，例如从新数据源添加一个特征，或使用更长的数据历史，都可能引入不同的法律依据、保留规则或公平性风险。他们的工具能够记录技术上的版本变化，却无法揭示这些监管后果。当团队需要说明为何模型中包含某些字段时，他们往往不得不手动还原来龙去脉。参与者 R2 将这一过程描述为“我们不该做的侦探工作”。

这些缺口在参与者的方法和文档实践中也显而易见。团队结合使用统计验证、分布监测、偏差测试和人在回路审查来评估数据质量，但他们指出，这些方法是围绕模型性能而非监管需求设计的。公平性评估通常依赖临时脚本，而非标准化流水线；没有组织表示拥有统一的工具来进行无偏差评估，以满足《人工智能法案》第10条的数据治理要求。文档工作流程同样支离破碎。尽管从业者提到了数据表、模型卡和元数据表，但这些材料通常都是事后编写的，与流水线执行脱节，而且很快就会过时。金融行业的受访者R2称，每年都要花数周时间编写一份200页的合规报告，将其技术工作流程转化为通俗易懂的语言。即使引入了生成式人工智能工具或内部聊天机器人，参与者（R5和R6）也表示，这些助手可以总结代码或元数据，却无法让文档与实际的机器学习流水线自动保持同步，团队因此不得不重新手动编辑。

总体而言，参与者的经历揭示了一种以工程性能为导向、却未能满足监管要求的工具生态。现有工具无法充分支持合规所要求的一致性、可追溯性、可解释性和文档记录。因此，从业者不得不依赖人工流程、临时脚本、领域直觉以及跨团队协商；随着监管审查日益严格，这些做法难以为继。

数据团队与法律团队之间的协作和责任缺口。参与者一再强调，他们面临的许多挑战并非纯粹的技术问题，而是源于数据团队与法律或合规团队之间的组织和沟通缺口。他们表示，要将《通用数据保护条例》或《人工智能法案》中的高层级监管原则转化为具体的运营数据质量实践，存在相当大的模糊性。具体任务的责任归属往往不明确。从业者 R6 这样描述这种困惑：

> “我们以为这些流程应该由法律团队审核。法律团队却以为是我们在负责管控。”

这些模糊不清的界限造成了不确定性和延误，尤其是在系统即将部署或审计临近时。

协作实践反映了这些结构性矛盾。参与者解释说，数据团队与法律团队之间的合作通常是非正式且被动的。双方都表示，希望能在流程的更早阶段参与，而不是直到最后才被征询意见。合规工作介入较晚时，技术团队便难以在时间压力下调整流水线设计、特征集或文档。正如从业者 R6 所说：“我们希望能更早参与合规程序。”法律团队也表达了类似的挫折感，并指出，尽早参与有助于他们提供更有价值的指导。在一些组织中，技术团队依靠内部合规倡导者来弥合这一差距，但这些人工作负担沉重，还要花费大量时间将监管指导转化为可执行的工程步骤。

这些叙述表明，在许多组织中，数据团队与法律团队之间的协作和责任边界仍未得到明确定义。缺乏共同语言、定期沟通或更明确的职责归属时，符合监管要求的数据质量很大程度上依赖跨团队协商和个人主动性。结果是一个

由临时协商推动，而非依托稳定、整合的工作流程的合规流程。

数据质量工作往往隐而不见、得不到认可、资源不足，而且总是被动应对。参与者认为，数据质量工作至关重要，却往往未受到足够重视，尤其是在涉及监管影响时。许多人将其描述为一种隐形劳动：必不可少，却很少纳入规划、得到认可或获得充足资源。因此，符合监管要求的质量实践往往是在审计或事件发生后才被发现并启动，而不是主动融入机器学习工作流程。

几位从业者指出，只有在出现故障或监管机构要求提供证据后，他们的团队才会尝试重建数据沿袭。工程师 R3 表示，他们的团队只会在"出了问题"时才绘制数据沿袭图，即便如此，整个过程仍需手动完成，而且并不完整。参与者还提到了安全性和访问控制方面的妥协。在交付压力下，团队有时会授予广泛的访问权限以加快开发，即使这与内部政策相冲突。审查周期和定期风险评估也很少开展。一些受访者表示，合规被视为一次性的障碍。法律团队批准初始设计后，团队便认为，在下一次重大审查之前，他们"可以随心所欲"（R11）。这种心态使得模型不断演进或引入新的数据源时，很难持续保障符合监管要求的数据质量。

在访谈中，参与者普遍认为，只有出现问题时，监管合规的数据质量工作才会受到关注。他们表示，这项工作需要投入时间、协调各方并付出努力，但这些付出很少在绩效评估或项目时间表中得到认可。如果缺乏专门的时间、资源和认可，监管合规的数据质量工作往往只能在问题出现后被动处理，而非作为机器学习实践中的常规环节。

4.2 从业者管理合规数据质量所需的支持

所需的技术支持。参与者表示，他们需要能够将工程工作与监管要求联系起来的工具。许多现有系统可以验证数据或追踪模型版本，却无法帮助团队理解其法律影响。从业者希望具备合规意识的工具能够直接融入机器学习工作流。他们希望数据沿袭系统不仅记录数据如何流动，还能记录数据获准流动的原因，包括处理目的、保留规则和数据源条件。还有几位参与者提到，需要能够跨数据管道运行的删除和保留机制，而非仅在存储层发挥作用。参与者表示，这有助于减少他们在应对审计或删除请求时目前不得不进行的手动"侦查工作"。公平性和偏差工具也是一项缺失。大家希望有标准化、可重复的公平性评估，而不是临时脚本。他们还认为，与数据管道保持同步的文档工具至关重要。他们希望平台能够及早揭示风险，并在数据集、特征或模型版本需要接受法律审查时发出明确信号。参与者 R1 总结这一需求时说：

> "我们需要能说明合规情况的工具，而不仅仅是讲技术的工具。"

所需的组织和流程层面的支持。从业者强调，仅有技术工具还不够。他们需要更清晰的治理结构和稳定的流程，以支持符合监管要求的数据质量管理。参与者呼吁明确各方角色及其责任，并建立定期审查周期。他们还希望为数据集审批、公平性评估和审计准备制定共享流程。一些人表示，希望有标准模板或检查清单，帮助他们完成受监管的步骤，尤其是在准备文档或评估新数据集时。他们希望能简单明了地了解监管机构的要求，以及这些要求如何与日常工程任务相联系。有人建议成立跨职能工作组，成员包括数据团队、法律团队和产品负责人。这些支持有助于团队理解合规的良好实践，以及何时需要请法律专家参与。

文化层面的支持与思维方式转变。 参与者表示，要实现符合监管要求的数据质量，也需要文化转变。许多人描述了一种思维模式：只有在被要求时才会关注数据质量与合规，而不是将其作为持续开展的实践。他们希望领导层认识到，质量工作需要时间，并应纳入项目规划。他们期待形成一种文化，让团队主动解决质量问题，而不是只在审计或事件发生时才作出反应。另有参与者指出，组织应奖励预防问题的工作，而不仅仅是交付新功能的工作。从业者还强调，技术和法律人员之间需要建立共同语言。他们认为，营造更加开放的文化，让团队及早并定期讨论监管风险，能够减少救火式应对，并使治理更具可持续性。由各团队尽早参与所支持的协作文化，被视为打造可持续且值得信赖的机器学习系统的关键。

5 讨论与建议

本研究考察了从业者如何理解并在日常工作中落实符合监管要求的数据质量。研究结果表明，从业者了解合规要求，却难以将高层次的监管原则转化为具体的工程实践。这些挑战不仅源于技术局限，也与组织结构、工作流程碎片化以及数据团队和法律团队之间协作不足有关。总体而言，这些发现有助于将数据质量重新界定为一种社会技术过程，它受到监管、治理和工程约束的共同影响。本节将讨论这些发现的更广泛意义，并阐述其对技术实践、组织结构和未来研究的启示。

将数据质量重新理解为一种社会技术实践。研究结果表明，受监管的机器学习环境中的数据质量超越了传统的工程理念，但仍以这些理念为基础。从业者通常从工程视角看待数据质量，重点关注性能、模式验证、分布检查或流水线稳定性。然而，监管将数据质量视为问责制、透明度和权利保护方面的问题。这种差异使人们难以确定如何在日常工作中满足合规要求。例如，法律团队认为准确性意味着个人记录正确无误，而工程师关注的是转换后的数据集中的统计模式。

在对数据最小化、相关性和无偏差的解读上，也存在类似的差异。这些发现表明，符合监管要求的数据质量并非数据的一种静态属性，而是一种社会技术实践，技术基础设施与监管预期在其中相互作用。数据质量实践需要明确兼顾技术和监管层面的解读，而不能想当然地认为现有的工程检查自然就能确保合规。

监管与技术之间的转译鸿沟。本研究的一项核心发现是，监管要求与工程工作流之间存在持续的鸿沟。《通用数据保护条例》和《人工智能法案》阐述了宽泛的原则，但从业者需要与具体数据管道步骤相关联、可付诸实践的指导。这种模糊性在访谈案例中尤为明显。从业者难以判断如何应对子群体性能差距、是否应使用溯源不明的数据集，或如何为监管审计准备证据。参与者描述了一些情形：即使是微小的模型变更，例如新增一个特征或扩展训练窗口，也会触发新的合规义务，而他们使用的工具无法体现这些义务。

问题并非缺乏合规意愿，而是缺少能够帮助团队结合具体情境解读监管要求的机制。从业者将这项工作描述为缓慢、不确定，而且往往令人倍感压力。在实践中，这种转译问题会导致延误、合规不完整，以及对模型更新或新数据源的犹豫。这表明，需要建立更系统的机制，明确将监管原则与工程制品及工作流转换联系起来。此类机制可以帮助团队在问题于审计或事件中暴露之前，推断变更带来的合规影响。

对技术基础设施和机器学习实践的启示。研究结果凸显了开发能够将合规要求融入机器学习工作流的技术解决方案的必要性。首先，现有工具支持验证、监控和版本管理，但很少记录追溯数据使用原因、应保留时长，以及模型变更时适用哪些义务所需的法律元数据。这一缺口导致应对审计请求时需要投入大量人工工作。合规感知基础设施应将法律元数据与数据资产和模型版本关联起来。同样，系统还应支持删除请求，并能将其传递至派生特征和历史数据集，因为其中往往残留个人数据痕迹。参与者希望有能够同时体现技术和法律层面的数据血缘工具。其次，应将文档更紧密地融入机器学习工作流。自动生成的文档应与不断演进的流水线保持同步。第三，参与者还呼吁建立更稳健、标准化的公平性评估流水线。能够识别和管理这些依赖关系的工具将减轻合规工作的运营负担。这些缺口表明，面向监管要求的数据质量需要新型机器学习基础设施，将法律背景整合其中，而不是将合规视为外部报告任务。

对组织结构和治理的启示。 单靠技术改进无法弥合合规鸿沟。我们的研究结果表明，组织结构对符合监管要求的数据质量起着关键作用。质量任务的责任归属往往不明确。数据团队认为法律团队会提供详细规则，而法律团队则期望工程师设计合规工作流程。这种鸿沟造成了延误、不确定性和合规风险。双方尽早参与被视为至关重要，但这种情况往往缺失。这与《通用数据保护条例》第25条第1款关于设计和默认数据保护的要求相冲突，因为该要求预设了尽早考虑合规问题。我们需要更明确的治理结构。正式建立跨职能审查周期，明确准确性检查、文档和公平性评估的职责归属，并制定共同的内部指南，有助于及早发现合规问题，减少最后一刻的返工。组织也可以指定合规倡导者，但要让这些职位发挥作用，必须为其提供充足的时间和结构性支持。改进入职培训和内部指南也会有所帮助。从业者表示，希望能简单明了地了解监管要求，以及如何将这些要求转化为技术任务。如果缺乏共同理解，合规工作就会缺乏一致性，而且往往只能被动应对。

建立积极主动的数据治理文化。另一项关键发现是，需要转变文化，朝着积极主动的治理方向发展。从业者指出，数据质量和合规工作往往不为人所见，也得不到认可。团队通常只有在发生事件或临近审计时才会处理相关问题。这些工作需要投入时间并进行协调，因此应纳入项目规划、资源分配和绩效评估。转向积极主动的治理，有助于认识到数据质量是可信机器学习开发的基础。领导层的支持被视为至关重要，这有助于将数据质量视为产品可靠性和用户信任的一部分，而不是官僚主义负担。更积极主动的文化可以减少对个人主动性的依赖，并为实现监管对齐创造更稳定、可持续的实践。

6 结论

本文从实证角度呈现了欧盟从业者在机器学习系统工作时对符合监管要求的数据质量的看法。研究结果并未将数据质量视为纯粹的技术属性，而是强调法规要求如何重塑数据管道、职位和组织中的日常数据工作。

在回答**研究子问题1**时，本研究展示了从业者如何在实践中解读并落实关键的法规适配的数据质量维度，以及这些解读如何受到监管要求的影响。进一步考察日常工作流程后，研究结果也回答了**研究子问题2**。我们展示了现有工具、方法和基础设施如何支持以性能为导向的数据质量观，而对文档记录、可追溯性和生命周期治理等监管义务的支持仍然有限。与此同时，从业者的陈述凸显出，他们往往通过人工变通方案和临时做法来应对这些局限。最后，关于**研究子问题3**，本研究阐明了技术团队与法律或合规团队之间的协作模式如何影响符合监管要求的数据质量能否得到有效落实，以及责任、沟通和时机方面仍存在哪些缺口。

通过凸显从业者视角，本研究揭示了为使机器学习系统符合不断演变的监管框架而必须开展、却往往不为人所见的协调、阐释与权衡工作。

这些见解凸显了未来研究的必要性：弥合法律意图与技术实现之间的差距，并通过合规感知基础设施、更清晰的治理模式以及持续的跨学科合作，为从业者提供支持。

生成式人工智能声明

作者使用 ChatGPT 进行语法检查和文体润色。

伦理考量

本研究遵循 ACM道德与职业行为准则。研究方案已获得作者所在机构伦理委员会的批准。参与完全自愿，并基于知情同意。鉴于合规相关主题的敏感性，所有个人及组织标识信息均已匿名化。

参考文献

[1] AI Index指导委员会。2025年。《2025年人工智能指数报告》。https://hai.stanford. edu/ai-index/2025-ai-index-report。访问日期：2025-02-05。[2] Apache软件基金会。2024年。Apache Atlas：面向 Hadoop 的数据治理与元数据框架。https://atlas.apache.org/ 软件可从 atlas.apache.org 获取。[3] Masoud Barati、Gagangeet Singh Aujla、Jose Tomas Llanos、Kwabena Adu Duodu、Omer Farooq Rana、Madeline Carr 和 Rajiv Ranjan。2021年。面向在线医疗健康的隐私感知云审计：用于验证《通用数据保护条例》合规性。《IEEE工业信息学汇刊》18（2021），4808–4819。https: //api.semanticscholar.org/CorpusID:237626884 [4] 2000 Christine Barter 和 Emma Renold。《我想给你讲个故事》：探索情境短文在儿童与青少年定性研究中的应用。《国际社会研究方法学杂志》3，4（2000），307–323。[5] David A. Basin、Søren Debois 和 Thomas T. Hildebrandt。2018年。基于处理目的并出于必要性：《GDPR下的合规》。载于《金融密码学》。https: //api.semanticscholar.org/CorpusID:158481534 [6] 2016Carlo Batini、Monica Scannapieco 等。数据与信息质量。沙姆，瑞士：施普林格国际出版社 63（2016）。[7] Reuben Binns。2018年。机器学习中的公平性：政治哲学带来的启示。载于《公平性、问责制与透明度会议》。PMLR，149–159。[8] Eric Breck、Shanqing Cai、Eric Nielsen、Michael Salib 和 D. Sculley。2017年。机器学习测试评分：评估机器学习生产就绪度并减少技术债务的标准。2017年IEEE大数据国际会议（大数据）（2017），1123–1132。https://api.semanticscholar.org/CorpusID:6244440 [9] Peter Buneman、Sanjeev Khanna 和 Wang Chiew Tan。2001年。为何以及何处：数据溯源的特征描述。载于数据库理论国际会议。https://api.semanticscholar.org/CorpusID:13791826 [10] 2016 Kelly E. Caine。CHI样本量的本地标准。2016年CHI 人类因素与计算系统会议论文集（2016）。https: // api.semanticscholar.org/CorpusID:14445824 [11] 2021 Kyla Chasalow 和 Karen E. C. Levy。统计学、政治学与机器学习中的代表性。2021年ACM 公平性、问责制与透明度会议论文集（2021）。https://api.semanticscholar. org/CorpusID:231573201 [12] James Cheney、Laura Chiticariu 和 Wang Chiew Tan。2009年。数据库中的溯源：为何、如何以及何处。《数据库基础与趋势》1（2009），379–474。https://api.semanticscholar.org/CorpusID:1778356 [13] 2022 Rishav Chourasia、Neil Shah 和 R. Shokri。摒弃遗忘：迈向机器学习中的真正数据删除。载于机器学习国际会议。https://api.semanticscholar.org/CorpusID:252917763 [14] 2021 Nicole M Deterding 和 Mary C Waters。深度访谈的灵活编码：二十一世纪的方法。《社会学方法与研究》50，2 2021 708 739（），–。[15] Sebastian Deterding。2011年。游戏元素所处情境下的动机可供性：概念模型。载于《游戏化：在非游戏情境中运用游戏设计元素》，CHI专题研讨会，第10卷。[16] Sebastian Deterding、Dan Dixon、Rilla Khaled 和 Lennart Nacke。2011年。从游戏设计元素到游戏性：定义"游戏化"。载于第15届国际学术MindTrek会议论文集：展望未来媒体环境（芬兰坦佩雷）（MindTrek '11）。计算机协会，美国纽约州纽约市，9–15。doi:10.1145/2181037.2181040 [17] Ying Dong、Wenbo Guo、Yueqi Chen、Xinyu Xing、Yuqing Zhang 和 Gang Wang。2019年。《面向公共安全漏洞报告中不一致性的检测》。载于 USENIX安全研讨会。https://api.semanticscholar.org/ CorpusID:148567270 [18] Lilian Edwards 和 Michael Veale。2017年。《算法的奴隶？为什么"解释权"可能并非你所寻求的补救措施》。《杜克法律与技术评论》 16 （ 2017 年），18–84。https://api.semanticscholar.org/CorpusID: 59148202[19] Lilian Edwards 和 Michael Veale。2018年。《奴役算法：从"解释权"到"获得更好决策的权利"？》《IEEE安全与隐私》16（ 2018 46 54 4049746），–。[20] 2016 https://api.semanticscholar.org/CorpusID: 欧洲议会和欧盟理事会。。欧盟条例（EU）2016/679：欧洲议会和理事会于2016年4月27日颁布的关于在处理个人数据及此类数据自由流动方面保护自然人的条例（《通用数据保护条例》）。http://data.europa.eu/eli/reg/2016/679/oj《欧盟官方公报》，L 119，2016年5月4日，第1–88页。[21] 欧洲议会和欧盟理事会。2024年。《欧洲议会和理事会关于制定人工智能统一规则（《人工智能法案》）并修订若干欧盟立法法案的条例》。http://data.europa.eu/eli/reg/2024/1689/oj [22] Rui Fu、Yuncheng Wu、Quanqing Xu 和 Meihui Zhang。2023 年。《 FEAST：面向关系数据的通信高效联邦特征选择框架》。ACM《数据管理论文集》1（2023年），1–28。https: // api.semanticscholar.org/CorpusID:259077221 [23] Marta Cantero Gamito 和 Christopher T Marsden。2024年。《人工智能共同监管？标准在欧盟《人工智能法案》中的作用》。《国际法律与信息技术杂志》32（2024年）。https://api.semanticscholar.org/CorpusID:271069058[24] Timnit Gebru、Jamie Morgenstern、Briana Vecchione、Jennifer Wortman Vaughan、Hanna M. Wallach、Hal Daumé 和 Kate Crawford。2018年。《数据集数据表》。《ACM通讯》64（2018年），86–92。https://api.semanticscholar.org/CorpusID: 4421027[25] Marissa Kumar Gerchick、Ro Encarnación、Cole Tanigawa-Lau、Lena Arm- strong、Ana Gutiérrez 和 Danaé Metaxa。2025年。《审计审计工作：从第144号地方法的偏见审计中汲取算法问责经验》。载于2025年 ACM 公平性、问责制与透明度会议论文集（FAccT '25）。计算机协会，美国纽约州纽约市，29–44。doi:10.1145/3715275.3732004[26] Antonio A. Ginart、Melody Y. Guan、Gregory Valiant 和 James Y. Zou。2019年。《让人工智能忘记你：机器学习中的数据删除》。ArXiv abs/1907.05012（ 2019 195886255 ）。[27] 2024 https://api.semanticscholar.org/CorpusID: Joan Giner-Miguelez、Abel G'omez 和 J ordi Cabot。。《使用大语言模型丰富机器学习数据集文档》。ArXiv abs/2404 15320 2024 269330239 。（）。[28] 2017 https://api.semanticscholar.org/CorpusID: Michelle Goddard。。《欧盟《通用数据保护条例》（GDPR）：具有全球影响的欧洲法规》。《国际市场研究杂志》59, 6（2017年），703–705。[29] Abe Gong、James Campbell 和 Great Expectations。 [n. d.]。Great Expectations。https://github.com/great-expectations/great_expectations [30] Mark Grover 等。2019年。《Amundsen：Lyft 的数据发现与元数据引擎》。https://github.com/amundsen-io/amundsen Lyft工程部。[31] Amy Kathleen Heger、Elizabeth B. Marquis、Mihaela Vorvoreanu、Hanna M. Wal- lach 和 Jennifer Wortman Vaughan。2022年。《了解机器学习从业者对数据文档的看法、需求、挑战与期望》。ACM《人机交互》6（2022年），1–29。https://api.semanticscholar.org/CorpusID:249431472 [32] Melanie Herschel、Ralf Diestelkämper 和 Houssem Ben Lahmar。2017年。《溯源调查：用途为何？形式如何？来源何处？》《VLDB期刊》26（2017年），881–906。https://api.semanticscholar.org/CorpusID:32007099[33] Kenneth Holstein、Jennifer Wortman Vaughan、Hal Daumé、Miroslav Dudík 和 Hanna M. Wallach。2018年。《改善机器学习系统中的公平性：行业从业者需要什么？》2019年 CHI 人类因素与计算系统会议论文集（2018年）。https://api.semanticscholar.org/CorpusID: 54895308[34] Chris Jay Hoofnagle、Bart Van Der Sloot 和 Frederik Zuiderveen Borgesius。2019年。《欧盟《通用数据保护条例》：内容及其意义》。《信息与通信技术法》28, 1（2019年），65–98。[35] Aspen Hopkins、Isabella Struckman、Kevin Klyman 和 Susan S. Silbey。2025年。《救济、修复、赔偿与预防：人工智能供应链的利益相关者分析》。载于2025年 ACM 公平性、问责制与透明度会议论文集（FAccT '25）。计算机协会，美国纽约州纽约市，209 227 10 1145 3715275 3732017 –。doi: . / . Zachary Izzo、Mary Anne Smart、Kamalika Chaudhuri 和 James Y. Zou。2020年。《从机器学习模型中近似删除数据：算法与评估》。ArXiv abs/2002.10077（2020年）。https://api.semanticscholar.org/ CorpusID:232033672 [37] Justin M. Johnson 和 Taghi M. Khoshgoftaar。2019年。《类别不平衡条件下深度学习综述》。《大数据期刊》6（2019年），1–54。https://api.semanticscholar.

102354936
org/CorpusID:[38] 2024 Kubeflow 作者。。《Kubeflow：面向Kubernetes 的机器学习工具包》。https://www.kubeflow.org/ 开源软件项目。[39] Steinar Kvale。2009年。《访谈：学习定性研究访谈的技巧》。Sage。[40] Clément Labadie 和 Christine Legner。2023年。《构建数据管理能力以应对数据保护法规：来自欧盟《通用数据保护条例》的经验》。《信息技术杂志》38, 1（2023年），16–44。[41] Mars Lan 等。2019年。《DataHub：在 LinkedIn 推广数据发现》。https: // datahubproject.io/ LinkedIn 工程团队。[42] Johann Laux、Sandra Wachter 和 Brent Mittelstadt。2023年。《可信人工智能与欧盟《人工智能法案》：论可信度与风险可接受性的混淆》。《监管与治理》18 (1): 3–32。[43] Johann Laux、Sandra Wachter 和 Brent Daniel Mittelstadt。2023年。《可信人工智能与欧盟《人工智能法案》：论可信度与风险可接受性的混淆》。《监管与治理》18（2023年），3–32。https://api.semanticscholar.org/CorpusID:
260498921
[44] Jonathan Lazar、Jinjuan Heidi Feng 和 Harry Hochheiser。2017年。《人机交互研究方法》。Morgan Kaufmann。[45] Michael Madaio、Shivani Kapania、Rida Qadri、Ding Wang、Andrew Zaldivar、Remi Denton 和 Lauren Wilcox。2024年。《在职学习负责任人工智能：学习路径、取向与抱负》。载于2024年 ACM 公平性、问责制与透明度会议论文集。1544–1558。[46] Michael A. Madaio、Luke Stark、Jennifer Wortman Vaughan 和 Hanna M. Wallach。2020年。《协同设计检查清单，以了解组织在人工智能公平性方面的挑战与机遇》。2020年 CHI 人类因素与计算系统会议论文集（2020年）。https://api.semanticscholar.org/ CorpusID:210156214[47] Margaret Mitchell、Simone Wu、Andrew Zaldivar、Parker Barnes、Lucy Vasserman、Ben Hutchinson、Elena Spitzer、Inioluwa Deborah Raji 和 Timnit Gebru。2018年。《用于模型报告的模型卡》。公平性、问责制与透明度会议论文集（2018年）。https://api.semanticscholar.org/CorpusID: 52946140[48] Curtis Northcutt、Lu Jiang 和 Isaac Chuang。2021年。《置信学习：估计数据集标签的不确定性》。《人工智能研究杂志》70（2021年），1373–1411。[49] Claudio Novelli、Federico Casolari、Antonino Rotolo、Mariarosaria Taddeo 和 Luciano Floridi。2024年。《认真看待人工智能风险：人工智能法案的新评估模型》。《人工智能与社会》39, 5（2024年），2493–2497。[50] Yaniv Ovadia、Emily Fertig、Jie Jessie Ren、Zachary Nado、D. Sculley、Sebastian Nowozin、Joshua V. Dillon、Balaji Lakshminarayanan 和 Jasper Snoek。2019年。《你能信任模型的不确定性吗？评估数据集偏移下的预测不确定性》。载于神经信息处理系统大会。https: // api.semanticscholar.org/CorpusID:
174803437
[51] 2019 Samir Passi 和 Solon Barocas。。《问题构建与公平性》。公平性、问责制与透明度会议论文集（2019年）。https://api.semanticscholar.org/CorpusID:57759365[52] Giorgio Patrini、Alessandro Rozza、Aditya Krishna Menon、Richard Nock 和 Lizhen Qu。2016年。《使深度神经网络能够抵御标签噪声：损失校正方法》。2017年 IEEE 计算机视觉与模式识别会议（CVPR）（2016 年），2233–2241。https://api.semanticscholar.org/CorpusID: 16406313[53] Maria Priestley、Fionntán O’donnell 和 Elena Simperl。2023年。《机器学习开发流水线中重要的数据质量要求综述》。ACM《数据与信息质量杂志》15, 2（2023年），1–39。[54] Joaquin Quionero-Candela、Masashi Sugiyama、Anton Schwaighofer 和 Neil D. Lawrence。2009年。《机器学习中的数据集偏移》。https://api.semanticscholar. org/CorpusID:[55]
61294087
Ricardo Trainotti Rabonato 和 Lilian Berton。2024年。《机器学习公平性的系统性综述》。《人工智能与伦理》5（2024年），1943–1954。https://api. semanticscholar.org/CorpusID:272795996[56] Inioluwa Deborah Raji、Andrew Smart、Rebecca N White、Margaret Mitchell、Timnit Gebru、Ben Hutchinson、Jamila Smith-Loud、Daniel Theron 和 Parker Barnes。2020年。《弥合人工智能问责差距：定义端到端的内部算法审计框架》。载于2020年公平性、问责制与透明度会议论文集。33–44。[57] Bogdana Rakova、Jingying Yang、Henriette Cramer 和 Rumman Chowdhury。2020年。《负责任人工智能与现实相遇之处》。ACM《人机交互》5（2020年），1–23。https://api.semanticscholar.org/CorpusID: 219966320[58] Sandeep Rangineni。2023年。《机器学习开发流水线框架的数据质量要求分析》。《国际计算机趋势与技术杂志》71, 9（2023年），16–27。[59] Khushnaseeb Roshan 和 Aasim Zafar。2024年。《数据流中的集成自适应在线机器学习：网络入侵检测系统案例研究》。《国际信息技术杂志》16（2024年），5099–5112。https://api.semanticscholar.org/CorpusID:267713401

[60] Nithya Sambasivan、Shivani Kapania、Hannah Highfill、Diana Akrong、Praveen Paritosh 和 Lora M Aroyo。2021年。“人人都想做模型，而不想做数据”：高风险人工智能中的数据级联。载于《2021年 CHI 人类因素与计算系统会议论文集》。1–15。[61] Holli Sargeant 和 Måns Magnusson。2025年。自动化决策系统中反歧视法的形式化。载于《2025年 ACM 公平性、问责制与透明度会议论文集》。181–194。[62] Pegdwendé N. Sawadogo 和 Jérôme Darmont。2020年。论数据湖架构与元数据管理。《智能信息系统期刊》56（2020），97–120。https://api.semanticscholar.org/CorpusID:220064078[63] Sebastian Schelter、Felix Biessmann、Dustin Lange、Tammo Rukat、Phillipp Schmidt、Stephan Seufert、Pierre Brunelle 和 Andrey Taptunov。2019年。使用 Deequ 对数据进行单元测试。载于《2019年国际数据管理大会论文集》。1993–1996。[64] Sebastian Schelter、Dustin Lange、Philipp Schmidt、Meltem Celikel、Felix Biessmann 和 Andreas Grafberger。2018年。自动化大规模数据质量验证。《VLDB基金会会刊》11（2018），1781–1794。https://api.semanticscholar.org/CorpusID:
51989837
[65] David Sculley、Gary Holt、Daniel Golovin、Eugene Davydov、Todd Phillips、Dietmar Ebner、Vinay Chaudhary、Michael Young、Jean-Francois Crespo 和 Dan Dennison。2015年。机器学习系统中隐藏的技术债务。《神经信息处理系统进展》28（2015）。[66] Sean Sirur、Jason RC Nurse 和 Helena Webb。2018年。我们准备好了吗？理解遵守《通用数据保护条例》（GDPR）所面临的挑战。载于《第2届多媒体隐私与安全国际研讨会论文集》。88–95。[67] Jennifer Tridgell 和 Jatinder Singh。2025年。“受托管理”能否成为自由及开源软件治理中公平、可问责且透明的模式？超越欧盟《网络弹性法案》。载于《2025年 ACM 公平性、问责制与透明度会议论文集》（FAccT ’25）。计算机协会，
473 484 10 1145 3715275 3732032 –。doi: . / 。
[68] 2018 Michael Veale、Reuben Binns 和 Jef Ausloos。。当隐私保护设计与数据主体权利发生冲突时。《国际数据隐私法》8（2018），
105 123
58193452
–。
[69] 2021 https://api.semanticscholar.org/CorpusID: Michael Veale 和 Frederik J. Zuiderveen Borgesius。。解读欧盟《人工智能法案》草案——分析拟议方案中良好、不足及尚不明确的要素。《国际计算机法律评论》22（2021），
97 112
235765823
–。
[70] 2018 https://api.semanticscholar.org/CorpusID: Michael Veale 和 Lilian Edwards。。第29条工作组关于自动化决策与用户画像的指南草案：清晰之处、意外之处及待解问题。《计算机法律与安全评论》34（2018），398–404。https://api.semanticscholar.org/CorpusID:
47075884
[71] Michael Veale、Max Van Kleek 和 Reuben Binns。2018年。高风险公共部门决策中算法支持的公平性与问责制设计需求。载于《2018年 CHI 人类因素与计算系统会议论文集》。1–14。[72] Paul Voigt 和 Axel von dem Bussche。2024年。欧盟《通用数据保护条例》（GDPR）：实用指南。欧盟《通用数据保护条例》（GDPR）（
2024
273725799
）。
[73] 1996 https://api.semanticscholar.org/CorpusID: Richard Y Wang 和 Diane M Strong。。超越准确性：数据质量对数据消费者意味着什么。《管理信息系统期刊》12，4（1996），5–33。[74] Matei Zaharia 等。2018年。MLflow：用于管理机器学习生命周期的平台。https://mlflow.org/ [75] Matei A. Zaharia、Andrew Chen、Aaron Davidson、Ali Ghodsi、Sue Ann Hong、Andy Konwinski、Siddharth Murching、Tomas Nykodym、Paul Ogilvie、Mani Parkhe、Fen Xie 和 Corey Zumar。2018年。借助 MLflow 加速机器学习生命周期。《IEEE 数据工程公报》41（2018），39–45。https://api.
85459546
semanticscholar.org/CorpusID:[76] Daochen Zha、Zaid Pervaiz Bhat、Kwei-Herng Lai、Fan Yang、Zhimeng Jiang、Shaochen Zhong 和 Xia Hu。2025年。以数据为中心的人工智能：综述。《计算机调查》57，5（2025），1–42。

A 数据质量维度与法规映射

表 2 将数据工程文献中常用的数据质量维度映射到 GDPR 和《人工智能法案》中的相应要求。这些维度为分析符合监管要求的数据质量提供了结构化视角，并用于指导访谈方案的设计，包括问题提示和基于场景的短案例。

B 访谈方案与提纲

图2 展示了用于组织半结构式访谈的访谈提纲。

C 访谈案例

以下情境短文在访谈中用于引出关于符合法规的数据质量实践的讨论。这些情境均为假设，且刻意保持抽象，并不涉及参与者自己的组织或系统。

“有些人一时想不起合规方面的例子，所以我来给你一个简短的情境，看看你会如何处理。”

“请告诉我们，你们团队在当前环境下会如何切实处理这些情况。如果有些事情无法做到，或者你们需要额外支持，也请告诉我们，这些信息同样很有价值。”

“我们并不是在寻找‘完美’的答案，而是想了解你们团队在实践中会做出哪些可行选择，以及如何权衡。”

C.1 情境短文 1：偏差与子群体表现

“设想你已经部署了一个机器学习模型。整体性能看起来不错，但监控显示，该模型对某个特定子群体的表现明显较差，例如年长用户或某一特定性别群体。”

“在你们目前的工作环境中，你们团队会如何切实应对这种情况？”

“如果解决这个问题需要进行权衡，例如重新训练、收集新数据，或接受整体性能下降，我想了解你们会如何考虑这些决策。”

C.2 情境 2：数据集来源不明

“再给你们介绍一个情境。”

“你们团队收到一个可能提升模型准确性的第三方数据集，但相关文档并不完整。数据是如何收集的、处理目的是什么，以及是在什么条件下收集的，都不明确。”

“由于交付时间表的要求，管理层希望尽快使用这个数据集。”

“按照你们目前的配置，团队实际会如何应对这种情况？”

“如果有些事情无法做到，或者需要法律团队或其他团队提供额外支持，这些情况也有助于我们了解。”

C.3 情境 3：审计准备与文档

“现在想象一下，监管机构联系了你的组织。”

“他们要求你证明训练数据符合质量和合规要求。他们希望看到准确性检查、偏差评估，以及数据处理方式的文档证据。”

“你们的团队目前实际能够提供什么？”

“如果不投入额外工作，哪些信息会难以或无法展示？”

Recei 发布于2007年2月20日；修订于2009年3月12日；接受于5 Ju ne 2009 年

表 2：技术与监管视角（《通用数据保护条例》和《人工智能法案》）下数据质量维度的交集。

数据质量维度	技术解释	《通用数据保护条例》中适用的规定	《人工智能法案》中适用的规定
准确性	数据正确、可靠且不存在错误。	要求确保数据准确性；删除或更正不准确的个人数据——**第5条第1款(d)项，序言第39条** 更正不准确个人数据的权利——**第16条**	数据集应无错误 – **第10条第3款** 数据集应尽可能没有错误 – **序言第67条**
可追溯性	应对数据的变更或更新进行跟踪。	安全措施，包括日志– **第32条**	强制性的自动化日志记录和记录保存，以实现可追溯性– 性–**第12条第3款(b)和(c)项，第19条**
相关性	数据适合并适用于预期任务。	个人数据应当充分、相关，且仅限于处理目的所必需的范围 – **第5条第1款(c)项，序言第39条**	要求评估 数据集的适用性 – **第10条第2款(e)项** 数据集必须与预期目的相关 – **第10条第3款**
及时性	数据保持最新，并在需要时可供使用。	个人数据必须保持最新——**第5条第1款(d)项**	
完整性	数据的所有必要组成部分均已提供，并得到充分记录。	适用于预期目的的充分且相关的个人数据——**第5条第1款(c)项**	数据集必须完整，以实现预期目的——**第10条第3款** 完整性，以避免产生有偏差或不准确的人工智能输出 – **序言第67条**
文档	清晰、全面的信息应描述数据及其管理。	问责原则要求记录处理活动 –**第5条第2款，第35条，序言第82条** 要求保存处理活动记录 – **第30条第1款，(3) 和 (4)**	高风险人工智能系统应附有说明 关于输入数据的 – **第13条第3款第(b)项第(vi)目** 导致匹配的输入数据记录保存 – **第12条** 技术文档应包含以下信息：数据 –**第11条** ，**第18条**，**附件四** ，**附件十一**
可访问性	数据可由授权用户访问和使用获授权的用户。	数据主体的访问权 – **第15条** 以结构化格式提供数据可携带性（促进数据可访问性） – **第20条**	
安全性	个人数据应受到保护，避免未经授权的披露或滥用。	通过健全的保障措施确保完整性与保密性——**第5条第1款(f)项** 在系统开发中嵌入设计隐私/默认隐私——**第25条** 要求采取适当的技术和组织措施，确保安全水平与风险相适应——**第32条**	在人工智能生命周期内保护隐私，并遵循《通用数据保护条例》的原则——**序言第69条** 将隐私和数据保护纳入其风险管理人工智能系统的管理与伦理框架 –**序言 27** 《人工智能法案》不影响《通用数据保护条例》中的个人数据隐私 – **第2条第（7）款**
无偏见	数据公正且不存在偏见，并公平对待各群体。	自动化个人决策不应基于特殊类别的个人数据——**第22条第4款，序言第71段**	用于检测、预防和缓解可能偏见的数据治理和管理实践——**第10条第2款(f)和(g)项** 数据集应具有适当的统计属性，涉及个人或个人群体——**第10条第3款** 例外情况下处理特殊类别个人数据以进行偏见检测和纠正——**第10条第5款，序言第70条** 关注数据集中的偏见缓解——**序言第67条**

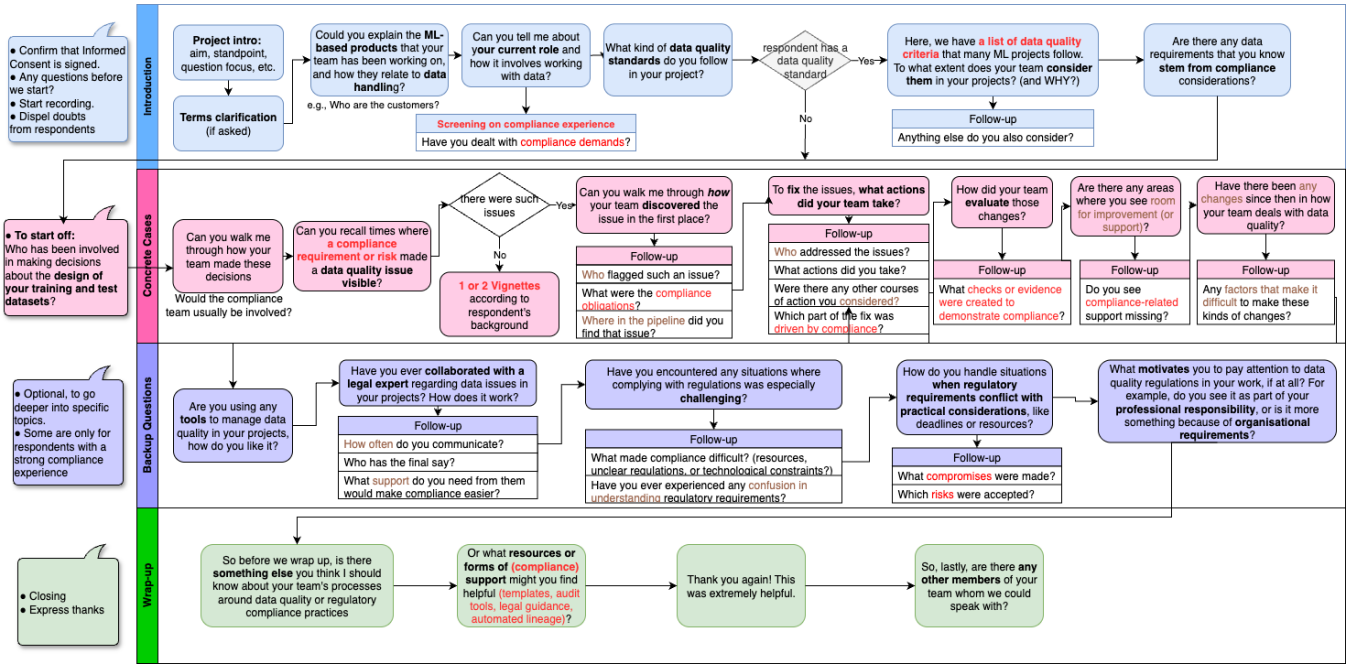


图2：半结构化访谈中使用的访谈提纲。该提纲指导了关于数据质量实践、合规相关挑战以及与监管要求相符的权衡的讨论。