arXiv:2602.04261v1 [cs.DB] 4 Feb 2026

数据智能体：级别、最新进展与开放问题

Yuyu Luo[†]香港科技大学（广州）广州，中国
y
uyuluo@hkust-gz.edu.cn

Guoliang Li
清华大学
北京，中国
liguoliang@tsinghua.edu.cn

Ju Fan
中国人民大学
中国北京
fanj@ruc.edu.cn

Nan Tang
香港科技大学（广州）
中国广州
nantang@hkust-gz.edu.cn

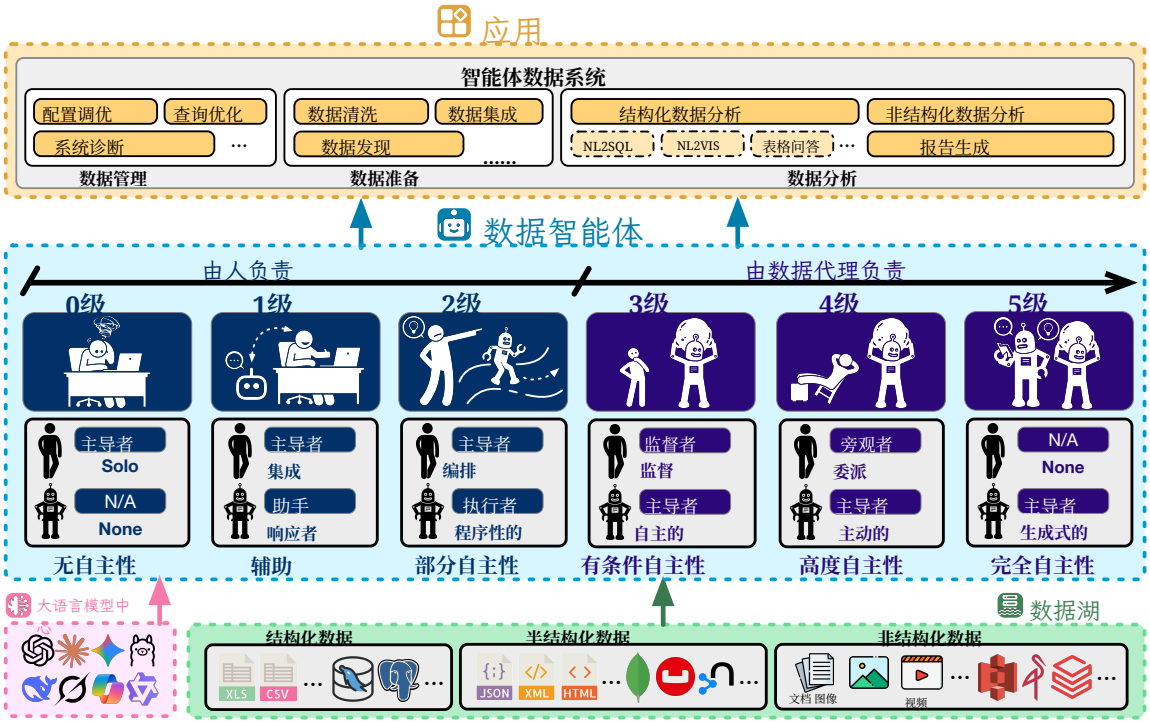


图 1：数据智能体概览 览。（https://github.com/HKUST Dial/awesome-data-agents）

摘要

数据智能体是一种新兴范式，利用大语言模型（LLM）和工具使用型智能体，自动执行数据管理、准备和分析任务。然而，“数据智能体”这一术语目前使用不一致，将简单的查询响应助手与理想化的完全自主“数据科学家”混为一谈。这种模糊性模糊了能力边界和责任归属，使用户、系统构建者和监管机构难以判断“数据智能体”能够以及不能够完成什么。

在本教程中，我们提出首个从 Level0（L0，无自主性）到 5级（L5，完全自主）的数据智能体分层分类体系。在此基础上，我们将介绍一种由生命周期和级别驱动的数据智能体视角。我们将（1）介绍 L0–L5 分类体系，以及区分简单助手与真正自主的数据智能体的关键演进飞跃；（2）回顾数据管理、准备和分析领域具有代表性的 L0–L2 系统；（3）重点介绍新兴的原型L3系统，这些系统致力于在监督下自主编排端到端数据工作流，以应对多样而全面的数据相关任务；（4）探讨面向主动的（L4）和生成式的（L5）数据智能体的前瞻性研究挑战。我们旨在为当今的系统提供一份实用指南，并为未来十年的数据智能体开发提供一份研究路线图。

特此授权，允许为个人或课堂使用而免费制作本作品的全部或部分数字或纸质副本，但前提是不得为营利或商业利益制作或分发副本，且副本首页必须附有本声明及完整引文。对于本作品中由非作者拥有版权的组成部分，必须尊重其版权。注明出处后进行摘要是允许的。除上述用途外，如需复制、再版、发布到服务器或重新分发至列表，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。*SIGMOD '26*，© 2026 版权所有归所有者/作者所有。出版权已授权给 ACM。ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM https://doi.org/XXXXXXX.XXXXXXX

1 引言

现代数据生态系统日益复杂，涉及异构和多模态数据源、不断演变的模式，以及紧密耦合的数据+人工智能流程 [33, 34, 39, 96]。与此同时，基于大语言模型的智能体在工具使用、规划和迭代推理方面展现出强大的能力 [40, 47, 50, 74, 89, 98]。因此，数据智能体一词迅速在学术界和工业界流行起来 [17, 45, 48, 68]，相关系统的范围从简单的 SQL 或商业智能聊天机器人，到以“完全自主的人工智能数据科学家”为卖点的雄心勃勃的产品。

†Yuyu Luo 为通讯作者。电子邮箱：yuyuluo@hkust-gz.edu.cn

然而，在缺乏共同术语的情况下，本质上截然不同的系统被混为一谈，统称为一个含义宽泛、被过度使用的术语。这会导致用户预期不符、发生故障时责任归属不明确，以及难以客观比较不同的方法。驾驶自动化领域此前也曾面临类似挑战，因此制定了 SAE J3016 标准，并提出了自主性的六级分类体系 [60]。

为了解决数据系统社区中的这一困惑，近期一项研究提出了数据智能体的分层分类体系 [100], ，从 0 级（L0，无自主性）到 5 级（L5，完全自主性），并沿着这一维度对现有系统进行了结构化综述，阐述了随着自主性提升，任务主导权和责任如何逐渐从人类操作人员转移到数据智能体。

在本教程中，我们基于该综述，将其转化为面向 SIGMOD 参会者的教学型框架。我们的目标是帮助参与者：（1）了解不同级别的数据智能体实际能够完成哪些任务；（2）梳理数据生命周期中不断扩展的系统版图；（3）确定推动数据智能体迈向更高自主性的关键研究挑战。

1.1 教程概述

我们将举办一场为期 3 小时的教程，其中包括 140 分钟的讲授（第 I–IV 部分），随后是 40 分钟的数据智能体实践（第 V 部分），供大家动手探索和讨论。

第一部分：问题定义与预备知识（30分钟）。我们首先阐述数据智能体在现代Data+AI生态系统中的重要性，并对数据智能体的概念进行形式化定义。我们将：（i）介绍数据智能体的动机与问题定义，强调现有“数据助手”系统为何不足，以及为何必须明确自主性与责任；（ii）更正式地定义数据智能体，并从环境、数据规模与结构、错误传播和治理要求等维度，将其与通用大语言模型智能体进行对比，并通过对比表突出这些差异；（iii）总结关键挑战（术语含义模糊、生命周期碎片化、自主性与治理之间的权衡、技术瓶颈），并说明为何需要建立数据智能体的分级分类体系。

第二部分：贯穿数据生命周期的 L0–L2级数据智能体（40 分钟）。接下来，我们聚焦较低的自主性级别（L0–L2级），并将其落实到数据生命周期的三个阶段：数据管理、数据准备和数据分析。我们将：（i）概述 L0、L1 和 L2 在每个阶段中的体现，并将其与图 1 所示的人类和智能体角色联系起来；（ii）深入每个阶段：在数据管理中，从人工数据库管理员（L0），到数据库调优/诊断/查询优化副驾驶（L1），再到可直接访问数据库管理系统与监控信号的 L2 智能体；在数据准备中，从脚本和规则（L0），到以建议方式开展数据清洗、集成和发现的副驾驶（L1），再到调用外部工具并通过执行反馈形成闭环的 L2 智能体；在数据分析中，从采用提示-响应范式的结构化数据分析（表格问答 / NL2SQL / NL2VIS）、非结构化数据分析和报告生成（L1），到能够感知环境、维持状态并调用 SQL、绘图和检索工具的 L2 分析智能体；（iii）使用一到两个贯穿示例（例如，数据库运维和 BI 分析），使 L0、L1 和 L2 之间的差异具体化。我们以此作结本部分，总结 L0–L2级反复出现的设计模式及其可靠性边界。

第三部分：L3数据代理与原型L3系统（45分钟）。接下来，我们将讨论3级这一当前研究前沿：数据智能体开始在人类监督下充当工作流编排器。我们将：（i）正式定义L3，并解释从L2到L3的关键演进飞跃；（ii）介绍学术界具有代表性的原型L3系统，这些系统探索了大语言模型编排器、语义算子、任务有向无环图优化和工具演进，以支持多样化的跨任务工作流，并讨论其架构、支持的任务、编排策略及局限性；（iii）分析云数据平台和湖仓中的工业级“数据智能体”产品，将其映射到相应级别，并强调常见的设计模式（例如，基于有向无环图的流水线编排、规划器与执行者分离、多智能体协作机制）及当前瓶颈（例如，依赖预定义算子/工具、因果与元推理能力有限、任务覆盖范围受限，以及对人工设计的防护机制的高度依赖）。

第四部分：迈向 L4–L5 与研究路线图（25 分钟）。最后，我们将通过探讨愿景中的第4级和第5级，并勾勒研究路线图，完成讲座部分。我们将：（i）阐述 L4数据代理的愿景：它们是主动的、长期运行且可自我治理的组件，能够持续监控数据+AI 生态系统，自主发现问题和机遇，并在没有明确指令的情况下编排流水线；（ii）介绍 L5数据代理，它们是生成式的数据科学家，能够构想新的解决方案、算法和范式，而不只是应用现有方法；（iii）总结关键开放问题，包括自主编排与多功能性、因果推理与元推理、内在动机与任务发现、长时域规划与权衡、安全与治理，以及自主性基准。

第五部分：数据代理实践平台——动手探索与讨论（40 分钟）。最后一部分是互动式的数据代理实践平台，旨在提升观众的参与度。我们将演示几个具体的数据代理工作流 [12, 20, 62, 66, 90], ，展示 L1/L2/Proto-L3 代理如何逐步运行，并邀请与会者试用我们自己的数据代理原型。我们将鼓励参与者为自己的场景构思或改进智能体，将其置于 L0–L5级谱系中，并讨论自主性、治理和可靠性方面的关键权衡；随后通过简短的问答，将这些洞察与第四部分的研究路线图联系起来。

1.2 我们的范围与目标

与现有教程的区别。关于大语言模型和数据系统的现有教程与综述通常聚焦于特定方面，例如用于数据库的大语言模型与数据分析 [33, 34, 41, 44, 73,95], 、机器学习数据管理 [8, 15, 96], ，或通用大语言模型智能体和工具使用系统 [66]。相比之下，我们的教程有以下三个特点：

(1) 基于级别的视角。我们采用基于级别的视角来审视数据智能体（L0–L5），明确关联自主性、能力与责任，从而更容易理解各级别“数据智能体”能够和不能够做什么。**(2) 整体生命周期视角**。我们采用整体生命周期视角，在统一的数据智能体框架下，全面涵盖数据管理、数据准备和

表1：通用大语言模型智能体与数据智能体的比较

方面	通用大语言模型智能体	数据智能体
主要重点	以任务和内容为中心： 完成既定任务或生成内容。	以数据生命周期为中心： 数据管理、准备和分析。
问题范围	自包含且静态： 根据明确指令和有限提示执行。	探索式且动态： 主动探索并浏览庞大的动态数据湖。
输入Data	小规模且即用型： 通常接收易于处理、干净的输入。	大规模且“原始”： 旨在处理异构、动态且含噪的原始数据。
Tool调用	通用工具包： 网络搜索、计算器、光学字符识别、图像生成器等。	专用数据工具包： 数据库加载器、*SQL* 等价性检查器、可视化库等。
主要输出	生成式产物： 人类可消费的产品：对话、推理、图像等。	数据产品与洞察： 配置、处理后的数据、洞察、可视化、分析报告等。
错误后果	本地化： 通常影响范围有限，仅涉及直接输出。	级联： 错误可能层层级联，进而影响下游洞见。

数据分析，而非将各项任务孤立看待。

(3) 演进飞跃与路线图。我们着重强调不同级别之间的演进飞跃，尤其是至关重要的 L2→L3 和 L3→L4 转变，并提出一份面向主动的（L4）和生成式的（L5）数据智能体的研究路线图，而非提供一份全面却平铺直叙的系统目录。

目标受众与学习成果。本教程面向广泛的 SIGMOD 受众，包括数据库、数据挖掘、机器学习、AI智能体和以数据为中心的 AI 领域的研究人员；构建数据平台、湖仓或企业数据技术栈的系统开发者和从业者；以及希望进入数据智能体这一新兴领域的学生。完成本教程后，参与者将能够使用 L0–L5框架对现有和未来的系统进行定位，区分数据智能体与通用大语言模型智能体，解读并评估供应商关于“数据智能体”的声明，为自己的应用选择合适的自主性级别，并理解感知、规划、工具、记忆和治理等关键设计维度。我们假设读者熟悉基本的数据库概念和大语言模型术语；教程内容本身将自包含。

2 教程大纲

我们首先定义数据智能体的含义，并将其置于数据系统和大语言模型智能体这一更广泛的领域中。

2.1 背景与问题定义

2.1.1 问题描述：什么是数据智能体？一般而言，数据智能体是一种基于大语言模型的架构，用于编排 Data+AI 生态系统，以执行配置调优、数据清洗、集成、探索和分析等数据相关任务 [17, 68, 95]。形式化地说，我们可以定义一个数据智能体 A，它在环境 E 中对原始数据 D 进行操作（例如，数据库管理系统、代码解释器、应用程序接口等），利用大语言模型 M，最终生成输出 O，以解决数据相关任务 T，抽象表示为：

$$\mathcal{A} : (\mathcal{T}, \mathcal{D}, \mathcal{E}, \mathcal{M}) \rightarrow \mathcal{O}.$$

这一宽泛的表述涵盖了各种系统，从能够建议 SQL 查询的简单助手，到能够自主管理和分析数据的理想化“人工智能数据科学家”。

2.1.2 任务版图与数据智能体和通用大语言模型智能体的比较。数据智能体运行于现代数据+人工智能生态系统中，这些系统涵盖关系型数据库、数据仓库和湖仓、数据湖，ETL/ELT管道、商业智能工具和机器学习服务。因此，数据智能体必须在不进行穷尽式摄取的情况下，对大型、异构且通常模式丰富的数据湖进行推理 [34]；与动态变化的噪声数据及系统交互，而这些系统的行为会随时间改变 [33]；并在多阶段管道中运行，其中错误可能悄无声息地传播并放大，而非仅影响单次响应。

因此，与通用大语言模型智能体相比，数据智能体面临的环境更加受限，但要求也显著更高。它们还需要满足严格的可靠性、治理和可复现性要求，而这些要求在许多通用智能体场景中并不那么突出。表1总结了数据智能体与通用大语言模型智能体在这些维度上的主要差异。

2.1.3 关键挑战。这些特征带来了若干挑战，也凸显了为数据智能体建立一套原则性分类体系的必要性：

- **术语含糊，且主张夸大**。由于缺乏通用词汇，自主性级别迥异的系统都被宣传为“数据智能体”，由此引发炒作、困惑和预期错位 [17, 68]。
- **数据生命周期各环节相互割裂**。数据智能体必须覆盖异构、多模态数据湖中的数据管理、数据准备和数据分析 [73, 95], ，但现有大多数研究都孤立地聚焦于单个任务或阶段，因此难以评估端到端能力并权衡利弊。
- **自主性与治理**。随着自主性提高，明确责任、界定安全运行范围并提供保障变得愈发重要，也愈发困难，尤其是在数据智能体能够自主修改数据、配置或流水线的情况下。
- **技术瓶颈**。要迈向更高的自主性级别，必须在感知（针对大型复杂数据和系统）、长时域规划与编排、记忆与持续适应、因果推理与元推理，以及与动态环境稳健交互等方面取得进展。

为明确这一点，我们采用一种面向数据智能体的基于等级的框架，明确关联自主性、能力和责任。

2.2 数据智能体的 L0–L5 层级

受 SAE J3016 驾驶自动化标准 [60],的启发，我们采用从 L0 到 L5 的数据智能体六级分类体系。如图 1 所示，数据智能体分为六个自主性级别，从 0级（L0）到 5级（L5）。图中指出了每个级别的数据相关任务由谁负责（人类与数据

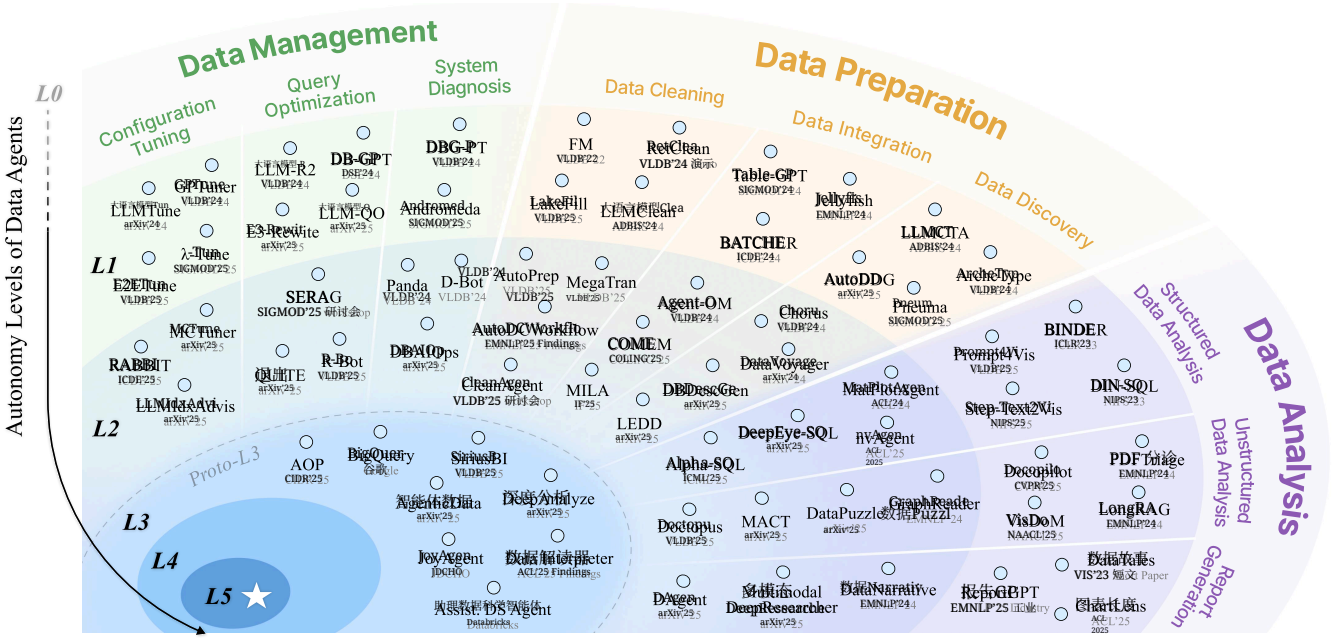


图 2：不同自主性级别的代表性数据智能体。

智能体），数据智能体扮演什么角色（例如响应者、执行者、编排者、主动的或生成式的组件），以及涉及数据生命周期的哪些环节（管理、准备、分析）。我们将在下文简要回顾这些级别。

2.2.1 自主性级别。

L0：无自主性。 在 L0 级别，数据智能体不参与任何任务。数据管理、准备和分析的所有任务均由人类手动完成。

L1：辅助。 L1 级别的数据智能体在无状态的提示—响应框架内运行。它们可以回答问题、生成代码片段或建议查询，但无法感知或与环境交互。人类仍全权负责执行和核验任何建议。

L2：部分自主性。 L2 级别的数据智能体能够感知并与其环境交互，包括数据湖、数据库管理系统、代码解释器和外部应用程序接口。它们可能具备记忆，并能够调用工具，在人工编排的流水线中自主执行特定任务流程。

L3：有条件的自主性。 L3数据代理预计能够在人工监督下，自主编排并执行针对各类任务定制的数据流水线。它们能够理解高层次的用户意图，并主导端到端工作流，而人类则充当监督者。

L4：高度自主性。 L4数据代理具备高度自主性和可靠性，无需人工监督或明确指令。它们被完全授权，可主动监控 Data+AI 生态系统，自主发现数据湖中的问题与机遇，并编排流水线以应对这些问题与机遇。

L5：完全自主性。 在 L5 阶段，数据智能体被设想为能够超越现有方法，创新出新的解决方案和范式，并作为完全自主且生成式的数据科学家开展工作。此时不再需要人类参与。

概括而言，图 2 展示了学术界和工业界的代表性系统在 L0–L5 各级以及数据生命周期的三个阶段中的位置。

2.3 L0–L2级：从人工工作流到部分自主性

本节将回顾 L0–L2级中具有代表性的系统，涵盖数据生命周期的三个阶段：数据管理、数据准备和数据分析。

2.3.1 数据管理。 数据管理包括数据库系统 [93, 96]中的配置调优、查询优化和系统诊断。在**L0**级，数据库管理员手动调整调优参数、索引配置和执行计划，依赖专业知识并反复试错 [93]。在 **L1**级，大语言模型被用作查询响应助手，以生成调优建议或重写后的查询。它们以提示—响应的方式工作，提供建议，由人类整合并验证 [18, 28, 38]。例如， *λ*-Tune [19]和 E2ETune [23] 使用大语言模型，根据工作负载特征推荐配置候选项；Andromeda [9] 则生成用于配置调试的诊断建议。在 **L2**级，数据智能体能够直接访问数据库管理系统和监控信息。它们可以观察工作负载统计信息、执行调优实验，并在决策循环中调整配置或重写查询，同时仍遵循人工设计的工作流 [63, 82, 94]。Rab- bit [67]、、R-Bot [69], 和 D-Bot [97]便是典型示例，它们利用环境反馈进行配置调优、查询重写和系统诊断。

2.3.2 数据准备。 数据准备 [7, 15, 32] 涵盖数据清洗 [4]、、集成 [35], 和发现 [16]。在 **L1** 阶段，数据智能体主要充当建议引擎：RetClean [51]和 LakeFill [83]推断并填补缺失值，LLMClean [4] 为清洗任务生成规则，Narayan 等人 [52]部署大语言模型来提出

模式匹配或实体对应，AutoDDG [88] 和 LLMCTA [27] 生成数据集摘要、元数据或列注释。同态压缩 [21] 是一种很有前景的方法，可以在保持语义完整性的同时降低数据智能体的计算成本。在 **L2** 阶段，数据智能体不再局限于查询响应器，而是直接与数据库或数据湖交互，以执行清洗和转换操作、验证约束，并根据执行反馈调整策略；随着探索更多数据，还会迭代优化集成决策 [26, 57, 91]。代表性系统包括用于数据清洗的 CleanAgent [57],和 MegaTran [32]；用于数据集成的 SEED [10], 和 Agent-OM [58]；以及用于数据发现的 LEDD [3]和 DBDescGen [37]。

2.3.3 数据分析。 数据分析包括结构化和非结构化数据分析，以及报告生成。在**L1**阶段，我们主要看到由大语言模型驱动的问答助手，用于表格问答 [11, 64, 86], NL2SQL [30, 42, 56, 99], NL2VIS [36, 43, 46, 49], 文本或多模态文档问答 [14, 59, 70, 79, 80],，它们针对精选数据集回答用户问题；此外还有报告生成器 [6, 65, 71]，用于处理输入表格或文档。在**L2**阶段，数据智能体不再局限于静态查询，而是能够动态参与多步骤分析流程，对其进行验证和完善 [55, 61, 75, 87] [29, 31], 。它们会调用 SQL引擎、绘图库 [53, 81, 85], 或检索模块 [25, 78, 92], 等工具，并支持对分析进行迭代探索和完善 [13, 54, 84]。

2.4 L3：迈向自主的数据智能体

现在我们转向3级（L3），这是从程序性执行者迈向自主编排器的关键一步。

2.4.1 从执行者到主导者。 在L2中，人类设计整体流水线，数据智能体则在这些由人类规定的工作流中执行特定程序。相比之下，在L3中，数据智能体需要理解高层次的用户意图，并自主编排涵盖数据管理、准备和分析的流水线。在执行过程中，数据智能体会根据反馈和中间结果调整流水线；人类则主要作为监督者，负责审查计划和结果，而非流水线设计者。从这个意义上说，任务主导权和主要责任从人类转移到了数据智能体。图3展示了典型的L3

数据智能体，突出了其在自主流程编排和优化方面的有条件自主性。

2.4.2 研究中的 *Proto-L3* 数据代理。近期研究系统开始展现出部分 L3 能力。它们使用基于大语言模型的编排器 [66], 预定义算子 [76, 77], 工作流优化 [22, 76], 和工具库 [90]，在异构系统上编排多步骤工作流，在单个智能体流程中覆盖数据生命周期的多个阶段，并在长期交互中保持状态，从而能够随着时间推移不断完善计划并纠正错误。这些 Proto-L3 数据代理通常在配备精选工具和数据的受限环境中运行，但它们为研究从侧重执行的 L2 智能体向以编排为中心的 L3 系统转变的过程提供了具体的测试平台。

我们将介绍若干具有代表性的学术系统，并讨论：(i) 它们的流程表示、编排和优化策略；(ii) 它们的架构选择（单智能体与多智能体、集中式与分布式规划器）；(iii) 它们在

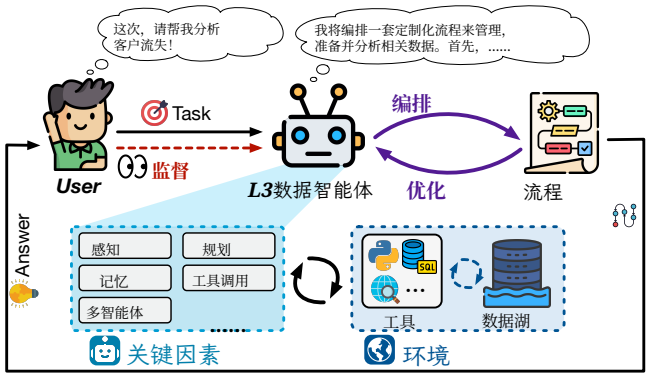


图 3：L3 数据智能体（有条件自主性）。

工具抽象与组合方面的做法；以及 (iv) 它们纳入反馈和处理错误的策略。

表 2 从工具灵活性、数据复杂性、数据生命周期覆盖范围，以及所支持的具体管理、准备和分析任务等维度，对比了学术界和工业产品中的代表性 Proto-L3 数据智能体。

2.4.3 工业数据智能体产品。工业平台（例如云数据仓库和湖仓）已开始推出商业化“数据智能体”产品 [20, 62]。我们分析：(i) 这些产品在实践中如何对应 L0–L3 级别；(ii) 它们提供哪些保障（例如人在回路确认、日志记录和回滚）；以及 (iii) 常见的局限性和设计权衡。

2.4.4 当前瓶颈与差距。该综述指出了若干阻碍当前系统实现完全 L3 自主性的差距：

- 流程编排能力有限，且依赖预定义算子；
- 高阶推理、因果推理和元推理能力不足，难以诊断并避免级联错误；
- 难以适应数据和工作负载不断变化的动态环境；
- 严重依赖人工设计的强化学习设置来实现对齐和落地。

这些挑战凸显了开发新方法的必要性，这些方法应超越简单调用工具的大语言模型智能体。

2.5 L4–L5：愿景与研究路线图

最后，我们将探讨富有远见的第4级和第5级，并概述一份研究路线图。

2.5.1 L4：主动型、高自主性数据智能体。在 L4 阶段，数据智能体被设想为数据+AI 生态系统中主动、长期运行且能够自我治理的组成部分。L4 智能体不再只是响应用户的明确请求，而是持续监控数据湖、系统和模型，检测数据漂移、性能退化和架构变更等现象，并发现有益的物化、缺失的索引或有前景的分析工作流等机会。它应能对这些任务进行优先级排序，自行设计和调整流程以应对这些任务，而无需明确指令；即使没有人工监督，也能在可靠性、安全和治理约束范围内运行。典型场景包括自主检测并缓解

表2：学术界研究与工业界产品中代表性 Proto-L3 数据代理的比较。比较开源：可用性；未预定义算子：利用未预定义算子的能力；数据相关任务覆盖范围涵盖数据管理、准备、分析；数据复杂性维度：多源（多用户）、异构（异构）和多模态（多模态）。

年份	数据智能体	开源-来源	未定义 Ops.	数据复杂性			数据管理			数据准备			数据分析		
				多用户	异构	多模态	配置 Tun.	查询 Opt.	Sys. 诊断	Data 清洗	Data 集成	Data 学科	结构化	非结构化	报告 Gen.
2025	AgenticData [66]	-	✗	✓	✓	-	-	✓	✓	✓	✓	✓	✓	✓	-
2025	DeepAnalyze [90]	✓	-	✓	✓	-	-	-	-	✓	✓	✓	✓	✓	✓
2025	AOP [76]	-	-	✓	✓	✓	-	✓	✓	✓	✓	✓	✓	✓	-
2025	iDataLake [77]	✓	-	✓	✓	✓	-	✓	-	✓	✓	✓	✓	✓	✓
2024	数据解释器 [22]	✓	-	-	✓	✓	-	-	-	✓	-	✓	✓	✓	✓
2025	JoyAgent [2]	✗	✗	✓	✓	-	-	-	✓	✓	✓	✓	✓	✓	✓
2025	助理 DS 代理 [12]	-	-	✓	✓	-	-	✓	-	✓	✓	✓	✓	✓	✓
2025	TabTab [72]	-	-	✓	✓	-	-	-	-	✓	✓	-	✓	✓	✓
2025	字节跳动数据智能体 [5]	-	-	✓	✓	-	-	-	-	-	✓	-	✓	✓	✓
2025	BigQuery [20]	-	-	✓	✓	-	-	✓	-	✓	✓	✓	✓	-	-
2025	皮质 [62]	-	-	✓	✓	✓	-	-	-	✓	✓	✓	✓	✓	-
2025	Xata Agent [1]	-	-	✓	✓	-	✓	✓	✓	-	-	✓	-	-	-
2025	SiriusBI [24]	-	-	-	-	-	-	-	-	✓	-	✓	✓	-	✓

工作负载变化、对索引和物化视图进行长时域管理，以及为关键数据资产持续开展质量保证。实现此类能力不仅会引发关于如何在整个数据生命周期中实现自主编排的问题，也需要建立相应机制，以支持内在动机、大型数据生态系统中的任务发现，以及能够权衡累计成本、延迟和数据质量取舍的长时域规划。

2.5.2 L5：生成式数据智能体。L5数据代理超越部署现有技术的范畴，被构想为自主的、生成式的数据科学家。一个L5数据代理应能识别当前方法中的缺口，在现有方法不足时对新的算法或表示提出假设，设计并分析实验以检验这些假设，并随时间迭代改进自身的解决方案。在这一愿景中，数据智能体不仅是数据库与机器学习系统的使用者，也是推动其演进的积极贡献者。迈向L5需要相应抽象，使数据智能体能够操控高层设计选择——例如物理设计、查询重写策略、数据清洗策略或学习过程——同时始终立足于可执行系统；还需要因果推理与元推理，以支持对备选设计进行有原则的诊断、比较与改进，乃至开创创新解决方案、新理论与新范式。

尽管完全实现的L4数据智能体与L5数据代理仍属推测，阐明这些层级有助于勾勒研究路线图。短期内，最紧迫的挑战在于使L2与Proto-L3智能体更加稳健、透明且可治理；中期内，迈向L4的进展将取决于自主的编排、任务发现以及多目标约束下长程决策方面的进展；而在更长期内，迈向L5将取决于将因果推理与元推理同由智能体驱动的实验相集成，并开发能够刻画自主性、适应性与安全、且超越传统任务级准确率的评估方法。

2.5.3 研究机遇。L0–L5层级揭示了若干与核心数据管理问题密切相关的研究方向。一个核心问题是，数据智能体应如何感知并操作大型异构数据湖：哪些索引、物化视图、摘要或学习得到的表示应充当它们的“感官”，这些结构应如何作为工具向其开放，以及智能体如何在管理、准备和分析过程中编排复杂的流水线，同时保障性能、数据质量和治理保障。

第二个主题关注如何在真实环境中训练和评估数据智能体。在这方面，运维日志、配置历史和遥测数据可以作为构建训练语料库、随时间调整智能体策略，以及支持针对故障与改进进行因果推理和元推理的基础。因此，我们需要制定基准测试和方法，不仅考察任务层面的准确率，还要评估数据智能体在真实数据管理工作负载中的自主性、稳健性、适应性和安全性。

3 个人简介

YuyuLuo是香港科技大学（广州）的助理教授，并在香港科技大学担任兼职职务。他于2023年获得清华大学博士学位。他的研究兴趣包括数据智能体、人工智能赋能数据库和以数据为中心的人工智能。他获得了2023年SIGMOD最佳论文。

Guoliang Li 是清华大学计算机科学系的正教授。他的研究兴趣主要包括数据清洗与集成，以及数据库机器学习。他曾获得VLDB 2017早期研究贡献奖、TCDE 2014早期职业生涯奖、VLDB 2023工业界最佳论文亚军、2023年SIGMOD最佳论文、SIGMOD 2023研究亮点奖、DASFAA 2023最佳论文奖和CIKM 2017最佳论文奖。

Ju Fan是中国教育部数据工程与知识工程实验室（DEKE Lab）和中国人民大学信息学院的教授。他于2013年获得清华大学博士学位，并获得ACM中国新星奖和2023年SIGMOD研究亮点奖。Fan博士的主要研究兴趣是人工智能赋能数据库和数据库系统。

Nan Tang 是香港科技大学（广州）副教授，并在香港科技大学兼任职务。他曾获VLDB 2010最佳论文奖、2023年SIGMOD研究亮点奖以及2023年SIGMOD最佳论文奖。他的主要研究兴趣是人工智能赋能数据库和以数据为中心的人工智能。

参考文献

[1] [暂无]。Xata Agent。[2] JDCHO 智能体团队。 [暂无]。JoyAgent-JDGenie。[3] 安琪、应驰华、朱雨晴、徐逸豪、张曼薇和王建民。2025年。LEDD：大语言模型赋能的数据湖数据发现。arXiv预印本arXiv:2502.15182（2025年）。[4] Fabian Biester、Mohamed Abdelaal 和 Daniel Del Gaudio。2024 年。LLM-Clean：通过大语言模型生成的开放式函数依赖实现上下文感知的表格数据清洗。arXiv: 2404 18681 [ ]。[5] [计算机科学数据库，字节跳动火山引擎。暂无]。数据智能体。[6] Lucas Cecchi 和 Petr Babkin。 [暂无]。ReportGPT：人在回路的可验证表格转文本生成。载于2024年自然语言处理实证方法会议：工业界专题论文集，Franck Dernoncourt、Daniel Preoţiuc-Pietro 和 Anastasia Shimorina（编）。[7] Chengliang Chai、Nan Tang、Ju Fan 和 Yuyu Luo。 [暂无]。揭秘数据准备中的人工智能。载于2023年国际数据管理会议伴随论文集。[8] Chengliang Chai、Jiayi Wang、Yuyu Luo、Zeping Niu 和 Guoliang Li。2023年。机器学习数据管理：综述。IEEE知识与数据工程汇刊（2023年）。[9] 陈思蓓、Ju Fan、吴斌、Nan Tang、邓超、王鹏翼、李晔、谭健、李飞飞、周景仁等。2025年。利用检索增强语言模型自动调试数据库配置。ACM数据管理会议论文集（2025年）。[10] 陈最、曹磊、Sam Madden、Tim Kraska、尚泽远、Ju Fan、Nan Tang、顾子晖、刘春伟和Michael Cafarella。2023年。SEED：利用大语言模型进行领域专用数据整理。arXiv预印本 arXiv:2310.00749（2023年）。[11] 程周军、谢天宝、史鹏、李承祖、Rahul Nadkarni、胡宇石、熊凯明、Dragomir Radev、Mari Ostendorf、Luke Zettlemoyer、Noah A. Smith 和余涛。2023年。在符号语言中绑定语言模型。arXiv: 2210 02875 [ ]。计算机科学, [12] [ ] Databricks。暂无。数据科学助手智能体。[13] 邓明航、Ashwin Ramachandran、徐灿文、胡兰香、姚哲伟、Anupam Datta 和张浩。 [暂无]。ReFoRCE：一种具备自我精炼、格式约束和列探索功能的文本转SQL智能体。arXiv预印本 arXiv:2502.00675（[暂无]）。[14] 段宇辰、陈哲、胡宇松、王维云、叶胜龙、史博天、陆乐维、侯启斌、卢同、李宏升、戴继锋和王文海。2025年。Docopilot：提升多模态模型的文档级理解能力。载于IEEE/CVF计算机视觉与模式识别会议论文集。[15] Alvaro AA Fernandes、Martin Koehler 等。2023年。数据准备：技术视角与综述。SN计算机科学（2023年）。[16] Benjamin Feuer、Yurong Liu、Chinmay Hegde 和 Juliana Freire。2024年。ArcheType：一种利用大语言模型进行开源列类型标注的新框架。VLDB会刊（2024年）。[17] 傅彦杰、王东杰、应王洋、张向良、刘欢和裴健。2025年。自主数据智能体：智能数据的新机遇。arXiv预印本arXiv:2509.18710（2025年）。[18] Victor Giannakouris 和 Immanuel Trummer。2024年。Dbg-pt：大语言模型辅助的查询性能回归调试器。VLDB会刊（2024年）。[19] Victor Giannakouris 和 Immanuel Trummer。2025年。 λ-tune：利用大语言模型实现数据库系统自动调优。ACM数据管理会议论文集（2025年）。[20] 谷歌云。 [暂无]。BigQuery。[21] 管家伟、张峰、马思齐、陈匡宇、胡一华、陈宇星、潘安群和Xiaoyong Du。2023年。同态压缩：实现压缩数据上的无限文本处理。ACM数据管理会议论文集1卷4期，第271篇文章（2023年12月），页码。doi：。/ Sirui Hong、林翊章等。Bang Liu。2025年。数据解释器：面向数据科学的大语言模型智能体。载于计算语言学协会论文集。[23] 黄欣梅、李浩洋、张静、赵欣欣、姚志明、李一言、张铁英、陈建军、陈红和李翠萍。2025年。E2Etune：通过微调生成式语言模型实现端到端参数调优。VLDB会刊（2025年）。[24] 姜杰、谢海宁、沈思齐、沈宇等。2025年。SiriusBI：面向商业智能数据分析的综合性大语言模型驱动解决方案。VLDB会刊（2025年）。[25] 蒋正宝、徐弗兰克、高露雨、孙志清、刘倩、Jane Dwivedi-Yu、杨一鸣、Jamie Callan 和 Graham Neubig。 [暂无]。主动检索增强生成。载于2023年自然语言处理实证方法会议论文集。[26] Moe Kayali、Anton Lykov、Ilias Fountalis、Nikolaos Vasiloglou、Dan Olteanu 和 Dan Suciu。2023年。CHORUS：用于统一数据发现的基础模型与探索。arXiv预印本arXiv:2306.09610（2023年）。[27] Keti Korini和Christian Bizer。2025年。面向基于大语言模型的列类型标注的知识生成与自我精炼策略评估。arXiv预印本arXiv:2503.02718（2025年）。[28] Jiale Lao、Yibo Wang、Yufei Li、Jianping Wang、Yunjia Zhang、Zhiyuan Cheng、Wanghu Chen、Mingjie Tang和Jianguo Wang。2024年。GPTuner：一种通过GPT引导的贝叶斯优化实现的人工阅读式数据库调优系统。VLDB会刊（2024年）。[29] Boyan Li、Chong Chen、Zhujun Xue、Yinan Mei和Yuyu Luo。2025年。DeepEye-SQL：一种受软件工程启发的文本转SQL框架。arXiv预印本arXiv:2510.17586（2025年）。[30] Boyan Li、Yuyu Luo、Chengliang Chai、Guoliang Li和Nan Tang。2024年。自然语言到SQL的黎明：我们是否已完全准备好？VLDB会刊（2024年）。[31] Boyan Li、Jiayi Zhang、Ju Fan、Yanwei Xu、Chong Chen、Nan Tang和Yuyu Luo。2025年。Alpha-SQL：使用蒙特卡洛树搜索的零样本文本转SQL。见第四十二届国际机器学习会议。[32] Changlun Li、Chenyu Yang、Yuyu Luo、Ju Fan和Nan Tang。2025年。面向高精度、低成本且可解释的数据转换的由弱到强提示与由轻量到强大的大语言模型。VLDB会刊（2025年）。[33] Guoliang Li、Jiayi Wang、Chenyang Zhang和Jiannan Wang。2025年。数据+AI：LLM4Data与Data4LLM。见2025年数据管理国际会议配套论文集。[34] Guoliang Li、Xuanhe Zhou和Xinyang Zhao。2024年。面向数据管理的大语言模型。VLDB会刊（2024年）。[35] Peng Li、Yeye He、Dror Yashar、Weiwei Cui、Song Ge、Haidong Zhang、Danielle Rifinski Fainman、Dongmei Zhang和Surajit Chaudhuri。2024年。Table-GPT：面向多样化表格任务的表格微调GPT。ACM数据管理会议论文集（2024年）。[36] Shuaimin Li、Xuanang Chen、Yuanfeng Song、Yunze Song、Chen Jason Zhang、Fei Hao和Lei Chen。2025年。Prompt4Vis：通过示例挖掘提示大语言模型进行表格数据可视化。VLDB期刊（2025年）。[37] Yuchen Li、Kai Wang、Zhiqiang Sun等。2025年。面向文本转SQL的数据库描述自动生成。arXiv预印本arXiv:2502.20657（2025年）。[38] Zhaodonghui Li、Haitao Yuan、Huiming Wang、Gao Cong和Lidong Bing。2024年。LLM-R2：一种用于提升查询效率的大语言模型增强型基于规则的重写系统。VLDB会刊（2024年）。[39] Xiaotian Lin、Yanlin Qi、Yizhang Zhu、Themis Palpanas、Chengliang Chai、Nan Tang和Yuyu Luo。2025年。LEAD：面向高效大语言模型指令微调的迭代式数据选择。CoRR（2025年）。[40] Bang Liu、Xinfeng Li、Jiayi Zhang、Jinlin Wang、Tanjin He、Sirui Hong、Hongzhang Liu、Shaokun Zhang、Kaitao Song、Kunlun Zhu等。2025年。基础智能体的进展与挑战：从脑启发智能到演化、协作与安全系统。arXiv预印本arXiv:2504.01990（2025年）。[41] Xinyu Liu、Shuyu Shen、Boyan Li、Peixian Ma、Runzhi Jiang、Yuxin Zhang、Ju Fan、Guoliang Li、Nan Tang和Yuyu Luo。2025年。大语言模型时代的文本转SQL综述：我们在哪里，又将去向何方？IEEE知识与数据工程汇刊（2025年）。[42] Xinyu Liu、Shuyu Shen、Boyan Li、Nan Tang和Yuyu Luo。2025年。Nl2sql-bugs：用于检测NL2SQL翻译中语义错误的基准测试。见第31届ACM SIGKDD知识发现与数据挖掘会议论文集第2卷。[43] Tianqi Luo、Chuhan Huang、Leixian Shen、Boyan Li、Shuyu Shen、Wei Zeng、Nan Tang和Yuyu Luo。2025年。nvBench 2.0：通过逐步推理解决文本转可视化中的歧义。arXiv: 2503 12880 [ ] . cs.CL [44] 2025 Yuyu Luo、Guoliang Li、Ju Fan、Chengliang Chai和Nan Tang。自然语言到SQL：最新进展与开放问题。VLDB会刊（2025年）。[45] Yuyu Luo、Xuedi Qin、Nan Tang和Guoliang Li。2018年。DeepEye：迈向自动数据可视化。见ICDE。IEEE计算机学会。[46] Yuyu Luo、Nan Tang、Guoliang Li、Chengliang Chai、Wenbo Li和Xuedi Qin。2021年。从NL2SQL基准测试合成自然语言到可视化（NL2VIS）基准测试。见2021年数据管理国际会议论文集（SIGMOD '21）。[47] Yuyu Luo、Nan Tang、Guoliang Li、Jiawei Tang、Chengliang Chai和Xuedi Qin。2022年。基于神经机器翻译的自然语言到可视化。IEEE可视化与计算机图形学汇刊（2022年）。[48] Yuyu Luo、Yihui Zhou、Nan Tang、Guoliang Li、Chengliang Chai和Leixian Shen。2023年。面向相似性搜索的折线图习得数据感知图像表示。ACM数据管理会议论文集（2023年）。[49] Paula Maddigan和Teo Susnjak。2023年。Chat2VIS：使用ChatGPT、Codex和GPT-3大语言模型通过自然语言生成数据可视化。IEEE Access（2023年10月）。[50] Shervin Minaee、Tomas Mikolov、Narjes Nikzad、Meysam Chenaghlu、Richard Socher、Xavier Amatriain和Jianfeng Gao。2024年。大语言模型：综述。arXiv预印本arXiv:2402.06196（2024年）。

[51] Zan Ahmad Naeem、Mohammad Shahmeer Ahmad、Mohamed Eltabakh、Mourad Ouzzani和Nan Tang。2024年。RetClean：使用大语言模型与数据湖的基于检索的数据清洗。VLDB会刊（2024年）。[52] Avanika Narayan、Ines Chami、Laurel Orr和Christopher Ré。2022年。基础模型能否整理你的数据？VLDB会刊（2022年）。[53] Geliang Ouyang、Jingyao Chen、Zhihe Nie、Yi Gui、Yao Wan、Hongyu Zhang和Dongping Chen。2025年。nvAgent：通过协作式智能体工作流实现基于自然语言的自动数据可视化。见第63届计算语言学协会年会论文集。[54] Fatemeh Pesaran Zadeh、Juyeon Kim、Jin-Hwa Kim和Gunhee Kim。 [无日期]。Text2Chart31：带自动反馈的图表生成指令微调。见2024年自然语言处理实证方法会议论文集。[55] Mohammadreza Pourreza、Hailong Li、Ruoxi Sun、Yeounoh Chung、Shayan Talaei、Gaurav Tarlok Kakkar、Yu Gan、Amin Saberi、Fatma Ozcan和Sercan O Arik。2025年。CHASE-SQL：文本转SQL中的多路径推理与偏好优化候选选择。见第十三届国际学习表征会议。[56] Mohammadreza Pourreza和Davood Rafiei。2023年。DIN-SQL：带自我纠正的文本转SQL分解式上下文学习。见神经信息处理系统进展第36卷，Alice Oh、Tristan Naumann、Amir Globerson、Kate Saenko、Moritz Hardt和Sergey Levine（编）。[57] Danrui Qi、Zhengjie Miao和Jiannan Wang。2025年。CleanAgent：使用基于大语言模型的智能体实现数据标准化自动化。arXiv: 2403 08291 。Zhangcheng [58] Qiang、Weiqing Wang和Kerry Taylor。Agent-OM：利用大语言模型智能体进行本体匹配。VLDB会刊（2024年）。[59] Jon Saad-Falcon、Joe Barrow、Alexa Siu、Ani Nenkova、Seunghyun Yoon、Ryan A. Rossi和Franck Dernoncourt。 [无日期]。PDFTriage：面向长篇结构化文档的问答。见2024年自然语言处理实证方法会议：工业界专题。[60] Elisabeth Shi、Tom Michael Gasser、Andre Seeck和Rico Auerswald。2020年。运行原理框架：面向自动驾驶功能的综合分类概念。SAE国际联网与自动驾驶汽车期刊（2020年）。[61] Zhihao Shuai、Boyan Li、Siyu Yan、Yuyu Luo和Weikai Yang。2025年。Deepvis：通过逐步推理连接自然语言与数据可视化。arXiv预印本 arXiv:2508.01700（2025年）。[62] Snowflake。 [无日期]。Cortex Agents。[63] Yuyang Song、Hanxu Yan、Jiale Lao、Yibo Wang、Yufei Li、Yuanchun Zhou、Jianguo Wang和Mingjie Tang。2025年。QUITE：一种超越规则、基于大语言模型智能体的查询重写系统。arXiv预印本arXiv:2506.07675（2025年）。[64] Yuan Sui、Mengyu Zhou、Mingjie Zhou、Shi Han和Dongmei Zhang。2024年。表格遇见大语言模型：大语言模型能否理解结构化表格数据？一项基准测试与实证研究。arXiv: 2305 13062 [ ] . cs Nicole [65] Sultanum和Arjun Srinivasan。DATATALES：探究使用大语言模型撰写数据驱动文章。见2023年IEEE可视化与视觉分析会议（VIS）。[66] Ji Sun、Guoliang Li、Peiyao Zhou、Yihui Ma、Jingzhe Xu和Yuan Li。2025年。AgenticData：面向异构数据的智能体数据分析系统。CoRR（2025年）。[67] Wenwen Sun、Zhicheng Pan、Zirui Hu、Yu Liu、Chengcheng Yang、Rong Zhang和Xuan Zhou。 [无日期]。Rabbit：检索增强生成实现更好的数据库参数自动调优。见2025年IEEE第41届国际数据工程会议（ICDE）。IEEE。[68] Zhaoyan Sun、Jiayi Wang、Xinyang Zhao、Jiachi Wang和Guoliang Li。2025年。数据智能体：一种用于编排数据+ AI生态系统的整体架构。arXiv预印本 arXiv:2507.01599（2025年）。[69] Zhaoyan Sun、Xuanhe Zhou、Guoliang Li、Xiang Yu、Jianhua Feng和Zhangm Yong。2025年。R-Bot：一种基于大语言模型的查询重写系统。VLDB会刊（2025年）。[70] Manan Suri、Puneet Mathur、Franck Dernoncourt、Kanika Goswami、Ryan A. Rossi和Dinesh Manocha。 [无日期]。VisDoM：使用多模态检索增强生成处理含视觉丰富元素的多文档问答。见2025年计算语言学协会美洲分会会议论文集：人类语言技术。[71] Manan Suri、Puneet Mathur、Nedim Lipka、Franck Dernoncourt、Ryan A. Rossi和Dinesh Manocha。2025年。ChartLens：图表中的细粒度视觉归因。见第63届计算语言学协会年会论文集。[72] TabTab AI。 [无日期]。TabTab。[73] Zirui Tang、Weizheng Wang、Zihang Zhou、Yang Jiao、Bangrui Xu、Boyu Niu、Xuanhe Zhou、Guoliang Li、Yeye He、Wei Zhou等。2025年。大语言模型/智能体即数据分析师：综述。arXiv预印本arXiv:2509.23988（2025年）。[74] Fengwei Teng、Zhaoyang Yu、Quan Shi、Jiayi Zhang、Chenglin Wu和Yuyu Luo。2025年。面向马尔可夫大语言模型测试时扩展的Atom of Thoughts。CoRR（2025年）。[75] Fen Wang、Bomiao Wang、Xueli Shu、Zhen Liu、Zekai Shao、Chao Liu 和 Siming Chen。2025年。ChartInsighter：一种借助基准数据集缓解时间序列图表摘要生成中幻觉问题的方法。arXiv: 2501 09349 [ ] 。计算机科学数据库 [76] Jiayi Wang 和 Guoliang Li。2025年。AOP：用于回答复杂查询的自动化交互式大语言模型流水线编排。CIDR。[77] Jiayi Wang、Guoliang Li 和 Jianhua Feng。2025年。iDataLake：一种面向数据湖的大语言模型驱动的分析系统。数据工程（2025年）。[78] Yuhao Wang、Ruiyang Ren、Junyi Li、Xin Zhao、Jing Liu 和 Ji-Rong Wen。 [无日期]。REAR：用于开放域问答的相关性感知检索增强框架。载于《2024年自然语言处理实证方法会议论文集》。[79] Yifan Wu、Lutao Yan、Leixian Shen、Yunhai Wang、Nan Tang 和 Yuyu Luo。2024年。ChartInsights：评估多模态大语言模型在低层次图表问答中的表现。载于 EMNLP（研究成果）。计算语言学协会，12174–12200。[80] Yupeng Xie、Yuyu Luo、Guoliang Li 和 Nan Tang。2024年。HAIChart：人机配对可视化系统。VLDB会刊（2024年）。[81] Wenyi Xu、Yuren Mao、Xiaolu Zhang、Chao Zhang、Xuemei Dong、Mengfei Zhang 和 Yunjun Gao。2025年。DAgent：由关系型数据库驱动的数据分析报告生成智能体。arXiv: 2503 13269 [ ] 。计算机科学数据库 [82] Zihan Yan、Rui Xi 和 Mengshu Hou。2025 无日期。MCTuner：通过大语言模型引导的探索和空间分解增强数据库调优。arXiv预印本 arXiv:2509.06298（2025年）。[83] Chenyu Yang、Yuyu Luo、Chuanxuan Cui、Ju Fan、Chengliang Chai 和 Nan Tang。2025年。利用数据湖，在数据冗余有限的情况下进行数据填补。VLDB会刊（2025年）。doi:10.14778/3748191.3748200[84] Zhaorui Yang、Bo Pan、Han Wang、Yiyao Wang、Xingyu Liu、Minfeng Zhu、Bo Zhang 和 Wei Chen。2025年。Multimodal DeepResearcher：利用智能体框架从零生成文本与图表交错报告。arXiv: 2506 02454 [ ] 。计算机科学语言学 Zhiyu Yang、Zihan [85] Zhou、Shuo Wang、Xin Cong、Xu Han、Yukun Yan、Zheng- hao Liu、Zhixing Tan、Pengyuan Liu、Dong Yu、Zhiyuan Liu、Xiaodong Shi 和 Maosong Sun。 [无日期]。MatPlotAgent：基于智能体的科学数据可视化方法与评估。载于《2024年计算语言学协会研究成果》。[86] Yunhu Ye、Binyuan Hui、Min Yang、Binhua Li、Fei Huang 和 Yongbin Li。2023年。大语言模型是多才多艺的分解器：为基于表格的推理分解证据与问题。载于《第46届国际ACM SIGIR信息检索研究与发展会议论文集》。[87] Peiying Yu、Guoxin Chen 和 Jingjing Wang。2025年。Table-Critic：用于表格推理协同评议与改进的多智能体框架。载于《第63届计算语言学协会年会论文集》。[88] Haoxiang Zhang、Yurong Liu、Aécio Santos、Juliana Freire 等。2025年。AutoDDG：使用大语言模型自动生成数据集描述。arXiv预印本 arXiv:2502.01050（2025年）。[89] Jiayi Zhang、Jinyu Xiang、Zhaoyang Yu、Fengwei Teng、Xionghui Chen、Ji- aqi Chen、Mingchen Zhuge、Xin Cheng、Sirui Hong、Jinlin Wang 等。2024年。Aflow：自动生成智能体工作流。arXiv预印本 arXiv:2410.10762（2024年）。[90] Shaolei Zhang、Ju Fan、Meihao Fan、Guoliang Li 和 Xiaoyong Du。2025年。DeepAnalyze：面向自主数据科学的智能体大语言模型。arXiv: 2510 16872 [ ] 。计算机科学人工智能 Yunfan [91] Zhang、Changlun Li、Yuyu Luo 和 Nan Tang。2024年。Sketchfill：用于填补派生缺失值的草图引导代码生成。arXiv预印本 arXiv:2412.19113（2024年）。[92] Zhengxuan Zhang、Zhuowen Liang、Yin Wu、Teng Lin、Yuyu Luo 和 Nan Tang。2025年。DataPuzzle：摆脱大语言模型在数据分析中幻觉式承诺的束缚。arXiv预印本 arXiv:2504.10036（2025年）。[93] Xinyang Zhao、Xuanhe Zhou 和 Guoliang Li。2023年。数据库参数自动调优：综述。IEEE知识与数据工程汇刊（2023年）。[94] Wei Zhou、Ji Sun、Xuanhe Zhou、Guoliang Li、Luyang Liu、Hao Wu 和 Tianyuan Wang。2025年。GaussMaster：基于大语言模型的数据库副驾驶系统。arXiv预印本 arXiv:2506.23322（2025年）。[95] Xuanhe Zhou、Junxuan He、Wei Zhou、Haodong Chen、Zirui Tang、Haoyu Zhao、Xin Tong、Guoliang Li、Youmin Chen、Jun Zhou 等。2025年。大语言模型数据综述。arXiv预印本 arXiv:2505.18458（2025年）。[96] Xuanhe Zhou、Guoliang Li 和 Zhiyuan Liu。2023年。大语言模型作为数据库管理员。arXiv预印本 arXiv:2308.05481（2023年）。[97] Xuanhe Zhou、Guoliang Li、Zhaoyan Sun、Zhiyuan Liu、Weize Chen、Jianming Wu、Jiesi Liu、Ruohang Feng 和 Guoyang Zeng。2024年。D-Bot：使用大语言模型的数据库诊断系统。VLDB会刊（2024年）。[98] Yizhang Zhu、Shiyin Du、Boyan Li、Yuyu Luo 和 Nan Tang。2024年。大语言模型擅长统计学吗？神经信息处理系统进展（2024年）。[99] Yizhang Zhu、Runzhi JIANG、Boyan Li、Nan Tang 和 Yuyu Luo。2025年。EllieSQL：通过复杂度感知路由实现成本高效的文本转SQL。载于第二届

语言建模会议.[100] Yizhang Zhu, Liangwei Wang, Chenyu Yang, Xiaotian Lin, Boyan Li, Wei Zhou, Xinyu Liu, Zhangyang Peng, Tianqi Luo, Yu Li, Chengliang Chai, Chong Chen, Shimin Di, Ju Fan, Ji Sun, Nan Tang, FugeeTsung, Jiannan Wang, Chenglin Wu, Yanwei Xu、Shaolei Zhang、Yong Zhang、Xuanhe Zhou、Guoliang Li 和 Yuyu Luo。2025年。数据智能体综述：新兴范式还是过度炒作？arXiv:2510.23587 [计算机科学数据库] https://arxiv.org/abs/2510.23587