面向智能体应用的数据盒重构与智能化生产框架

谷婧雯[1]，熊 [2,3]，朱扬勇[4]
1. 上海所迹数字科技有限公司，上海 200041；
2. 复旦大学计算与智能创新学院，上海 200438；
3. 上海市数据科学重点实验室，上海 200438；
4. 上海数据研究院有限公司，上海 200120

摘要

智能体应用对数据的任务适配性、业务语义表达及可调用性提出了新的要求，现有以数据质量控制为主的数据治理方法难以直接支撑智能体的任务执行。针对该问题，本文提出面向智能体应用的数据盒架构及智能化生产框架。首先，以任务需求为约束，将业务语义、业务规则及任务上下文引入数据组织过程，将数据盒由面向开放流通的数据封装单元扩展为面向智能体的数据组织单元。其次，提出数据母机，通过持续采集业务认知轨迹，协同更新业务模型、治理规则及数据盒配置，实现数据盒的持续构建、迭代更新及复用。该框架形成贯通业务认知、数据治理、数据组织及智能体调用的持续处理流程，为智能体应用中的数据组织及高质量数据集构建提供统一方法。

关键词

高质量数据集；智能体；数据盒；数据治理；数据组织

中图分类号：TP18；TP399 文献标志码：A doi：10.11959/j.issn.2096–0271.2026xxx

An Intelligent Production Framework for Data Box Reconstruction in Agent-Oriented Applications

Gu Jingwen[1], Xiong Yun[2,3], Zhu Yangyong[4]
1. Shanghai Suoji Digital Technology Co. , Ltd. ；
2. College of Computer Science and Artificial Intelligence, Fudan University；
3. Shanghai Key Laboratory of Data Science；
4. Shanghai Data Research Institute Co. , Ltd. , Shanghai 200120, China

Abstract

AI agents are emerging as an important form of large-scale AI deployment in domain-specific applications. However, existing data governance methods mainly focus on data quality, standardization, security, and management, and provide insufficient support for task-oriented data organization and machine-accessible business semantics required by agents. To address this problem,

this paper proposes a Data Box reconstruction method and an intelligent production framework for agent-oriented applications. First, business semantics, business rules, and task context are incorporated into the Data Box, extending it from a data encapsulation unit for open sharing and circulation to an agent-oriented data organization unit that integrates data content with business semantics. Second, a Data Machine is proposed to support the continuous construction of Data Boxes. It captures business cognition trajectories and coordinates business modeling with data governance, enabling iterative updates of business models, governance rules, and Data Box configurations. The proposed framework forms a closed-loop process from business cognition and data governance to Data Box construction and agent invocation, providing a systematic method for building industry-general and organization-specific datasets for agent-oriented applications.

Key words
high-quality dataset, AI agents, Data Box, data governance, data organization

1 引言

随着大模型能力持续提升，人工智能应用正由通用问答、内容生成等模式，拓展至面向具体业务任务的智能体应用。2025年，国务院印发《关于深入实施“人工智能+”行动的意见》，提出到2027年新一代智能终端、智能体等应用普及率超过70%。2026年政府工作报告进一步提出，促进智能体加快推广，推动重点行业领域人工智能商业化、规模化应用。智能体在垂直领域的规模化落地，对数据资源的任务适配性、业务语义表达和可调用性提出了更高要求。

智能体应用的任务性能不仅依赖模型能力，还依赖数据与业务任务之间的匹配程度。通用大模型能够提供跨领域的语言理解、推理和生成能力，但组织级任务通常绑定特定业务对象、业务流程、业务规则、权限约束和绩效目标。例如，设备故障诊断需要结合本组织的工艺参数、历史案例和处置规范，难以仅依赖通用训练语料完成。此类组织级专识具有领域专业性、上下文依赖性和组织专属性，通常分散于业务系统、文档、规则库及人员经验中，尚未被有效转化为可供智能体理解、检索和调用的数据资源。

现有数据治理主要关注数据标准、质量、安全和资产管理，为数据资源的规范化管理提供了基础。然而，其组织单元和治理过程通常缺少业务场景、任务目标及业务语义的显式牵引，治理后的数据难以直接满足智能体的任务理解、知识检索和工具调用需求。当数据治理与应用目标缺乏有效衔接时，容易出现治理范围与任务需求不匹配、数据语义表达不足以及治理价值难以评估等问题。因此，有必要从智能体应用需求出发，重新组织数据资源及其治理过程。

本文将面向智能体的数据资源划分为两个互补层次：一是支撑领域基础认知和通用任务能力的行业通识数据集；二是反映特定组织业务规则、流程、经验和场景知识的组织级专识数据集。前者为跨组织的领域智能体提供共享数据基础，后者则将组织内部数据和业务知识转化为面向特定任务的专属能力。两类数据资源共同构成智能体进入垂直领域的数据基础。

针对上述问题，本文提出面向智能体应用的数据盒架构与构建方法。首先，提

出以任务需求为导向的数据治理思路，将业务语义纳入数据组织过程；其次，以数据盒（Data Box）为基本组织单元，将其由支持开放流通的数据封装单元扩展为融合数据内容与业务语义的智能体数据组织单元；最后，提出数据母机（Data Machine）作为数据盒构建框架，通过持续采集业务认知轨迹，协同更新业务模型、治理规则和数据盒配置，实现数据盒的持续构建、更新与复用。本文形成贯通业务认知、数据治理、数据组织和智能体调用的闭环流程，为行业通识数据集和组织级专识数据集的建设提供方法参考。

2 相关工作

围绕智能体应用所需的数据基础，相关研究已从数据质量、语义组织、知识调用及数据产品等角度形成一定的成果，但如何从具体业务需求出发，将数据治理、业务语义及智能体任务执行贯通起来，仍有进一步研究空间。

2.1 高质量数据集及数据治理

在高质量数据集及数据治理方面，相关研究已由模型逐步转向数据，关注训练数据开发、数据调试、数据筛选及数据维护等问题。Gadre等研究不同多模态数据筛选策略[1]，Li等将数据选择扩展至语言模型预训练场景[2]。Shumailov等指出，递归使用生成数据可能导致模型性能退化[3]。面向人工智能的数据治理研究进一步强调数据质量、安全合规及全生命周期管理。李继峰等提出面向人工智能的数据治理框架[4]；陈璐等利用联邦学习开展跨源数据质量检测[5]；Jarrahi等强调数据工作应贯穿人工智能生命周期[6]；Yan等利用混合专家方法集成错误检测、清洗、补全及实体匹配等任务[7]。这些工作提升了人工智能系统的数据基础质量，但总体仍以模型训练及数据质量控制为主要目标，对于智能体在具体业务任务中需要何种数据、治理后的数据如何直接支撑任务执行，研究仍相对不足。

2.2 知识组织及智能体知识

知识库、本体及知识图谱被广泛用于异构数据的语义表示及关系组织。知识库、记忆及工具调用逐渐成为智能体任务执行的重要支撑，部分研究进一步通过世界模型、知识图谱等方式描述环境状态及其演化关系。任磊等利用数字族谱支撑工业具身智能世界模型[8]；Anokhin等将语义记忆及情景记忆组织为知识图谱形式的世界模型[9]；Guan等利用大语言模型构建可验证的符号世界模型，并与规划器结合完成任务执行[10]。这些研究为智能体理解和利用结构化知识提供了基础，但对企业场景中业务对象、规则、流程及数据之间如何形成统一映射，以及业务语义如何进一步反向约束数据治理范围，讨论仍较有限。

2.3 数据产品化及数据盒

数据产品化研究为数据从资源管理走向标准化使用及流通提供了重要路径。数据盒最初面向数据自治开放提出，通过封装数据及自主程序单元，实现数据受控访问、权益保护及可追踪使用[11]。在此基础上，叶雅珍、朱扬勇提出“盒装数据”，利用“时间—空间—内容”三维结构描述盒内数据，并支持数据计量及计价[12]；熊赟、朱扬勇进一步围绕数据流通标的辨识、追溯、安全及监管构建数据产品流通体

系[13]。随后，吴俊等提出数据产品统一标识编码方法[14]，杨琳、朱扬勇提出高质量数据集产品的“五元组形态”及生产流程[15]。这些研究推动数据建设由资源管理逐步走向标准化、产品化及可复用生产。但现有数据盒及数据产品研究主要关注数据封装、辨识、计量、流通及管理，对智能体任务执行所需的业务对象、业务规则及任务上下文表达仍显不足。

综上，现有研究已经在数据质量治理、知识结构化及数据产品化方面形成较为系统的方法，但面向智能体应用仍存在两个衔接问题：一是缺少以具体业务任务为牵引的数据组织方式，业务需求尚未充分参与治理对象及治理边界的确定；二是现有数据产品主要描述数据自身属性，对业务规则及任务上下文的表达不足。基于此，本文进一步扩展数据盒的业务语义表达能力，并提出数据母机的数据治理思路，将传统数据治理及业务建模结合起来，形成面向智能体应用的数据组织及生产机制。

3 面向智能体应用的数据盒重构方法

如第 2 节所述，数据盒主要面向数据开放、流通及监管场景。面对智能体的调用需求时，数据盒需要回答“这个数据单元对应哪个业务对象、能支撑什么操作、在什么规则下适用”等问题，需要由“面向开放的数据封装单元”进一步演化为“面向任务调用的数据组织单元”。

3.1 面向任务的数据—业务语义联合组织

智能体完成业务任务不仅依赖数据内容，还依赖数据所对应的业务对象、业务关系、业务逻辑及业务规则。例如，设备维修记录只有与设备、部件、故障现象及处置动作建立业明确关联，才能进一步支撑故障判断及处置决策。尽管大模型具备一定的语义表示能力，但在缺少与任务语义及业务约束的的情况下，仍难以稳定完成细粒度类型识别与任务适配 [16]。

为此，本文在原有数据盒基础上引入业务语义构件，在既定场景及任务目标约束下，对专家经验、业务规则、业务知识及经营逻辑等信息进行抽取与结构化表达，形成业务模型，并进一步与任务上下文关联。通过建立业务模型与数据对象之间的映射关系，使数据治理范围由面向一般数据资源的广泛治理，收敛为面向特定场景及任务需求的数据。

与原有数据盒相比，重构主要体现在两个方面。其一，在内部构成上，将业务语义纳入数据盒的基础构件，与数据内容、自主程序及权限约束共同构成面向智能体的四元数据组织结构。业务语义与数据内容建立显式关联，使数据单元能够同时携带业务含义、关系结构及适用规则，为智能体的数据理解与任务执行提供语义支撑。其二，在应用方式上，将数据盒与具体场景及任务目标绑定，由任务需求确定数据盒的治理边界与组织范围，使数据治理由一般性质量治理进一步转向面向应用的数据组织。

通过上述重构，数据盒在保留受控封装及调用能力的基础上，进一步形成面向特定场景与任务目标的数据组织载体，其质量评价也由侧重数据本身的完整性、准确性和一致性，扩展至数据对智能体任务调用的适配性与可靠性。

3.2 面向智能体调用的数据盒四元结构

面向智能体应用，数据盒由数据内容、

业务语义、自主程序及权限约束四类构件组成，分别承担数据供给、语义解释、能力调用及访问控制功能。

（1）数据内容。封装经筛选、清洗、标准化及关联处理后的结构化数据、文档等多源异构数据，其数据范围由具体业务场景及任务需求确定，为智能体执行任务提供基础数据支撑。

（2）业务语义。对业务对象、对象关系、业务规则、约束条件及任务上下文进行结构化表达，并与盒内数据建立显式映射，使智能体能够识别数据的业务含义及其在具体任务中的作用。

（3）自主程序。面向盒内数据提供查询、检索、统计、分析及模型调用等受控能力，并通过标准化接口向智能体开放，使上层应用能够在不直接访问底层数据的条件下完成数据调用与处理。

（4）权限约束。根据身份、角色、任务及数据对象设置访问边界，并记录数据调用及操作过程，实现数据使用的可控、可追踪及可撤销。

上述四类构件形成面向智能体的四元数据组织结构。其中，数据内容提供事实基础，业务语义建立数据与任务之间的语义关联，自主程序提供标准化调用能力，权限约束控制数据使用边界。四类构件协同作用，使数据盒能够在受控条件下支持智能体对业务数据的理解、检索、调用及任务执行。

3.3 面向应用的数据盒多层次价值表征方法

数据盒与具体应用场景及任务需求绑定后，可以形成边界明确、过程可追溯、调用可记录的数据计量单元。在此基础上，从治理成果、应用效能和经济价值三个层次构建数据盒价值表征方法，使数据治理产出能够由过程性活动进一步转化为可统计、可验证和可评价的对象。

（1）治理成果表征

数据盒具有明确的数据边界、业务范围、版本信息及治理记录，可作为数据治理成果的基本统计和审计单元。本文从封装完整率、语义关联率和版本可追溯率三个维度表征数据盒的治理完成程度。其中，封装完整率反映规定构件及说明信息的完备程度；语义关联率反映盒内数据与业务语义建立有效映射的程度；版本可追溯率反映数据内容、业务模型及治理过程的版本记录完整程度。上述指标所需信息可主要由数据盒封装过程中形成的元数据、语义映射及版本记录获取，从而降低额外评价数据的采集成本。具体计算方法如下：

封装完整率＝

说明文档四要素2实际覆盖项数÷4

语义调用率＝调用中无需人工解释即可被正确识别的业务对象、关系与规则字段数÷

盒内语义字段总数

版本可追溯率＝治理规则与数据映射关系的变更记录完整覆盖的版本数÷实际版本数

（2）应用效能表征

治理成果达到基本要求后，进一步从智能体实际调用过程评价数据盒的应用效能。本文选取有效调用频次、跨任务复用度和任务完成率等指标，分别刻画数据盒的实际使用程度、跨任务支撑能力及对任务执行的有效性。相关指标可依据数据盒接口调用日志、任务标识及执行结果计算，

使数据评价由数据本身的质量属性进一步延伸至其对智能体任务执行的支撑效果。具体计算方法如下：

有效调用频次＝
排除重复请求与测试调用后的实际调用次数

跨任务复用数＝调用该数据盒的不同任务数

任务完成率＝
调用后智能体任务成功完成的次数÷
有效调用次数

（3）经济价值表征。

在治理成果可统计、应用效能可验证的基础上，可进一步结合数据盒的建设成本、实际使用及业务收益开展经济价值评价。具体可采用成本法、收益法及市场法等方法。成本法以数据治理、标注、业务语义建模、封装及维护等直接投入为基础形成成本价值；收益法结合数据盒调用情况及其支撑业务产生的增量收益或成本节约，对预期经济利益进行估计；市场法则参考同类型数据产品的挂牌、交易或可比案例形成市场参照。不同方法适用于不同应用条件，可结合数据盒类型、使用方式及可获得证据综合选择。

上述三个层次构成递进的数据盒价值表征体系。治理成果表征回答数据盒“是否形成规范的数据成果”，应用效能表征回答数据盒“是否能够有效支撑智能体任务”，经济价值表征进一步分析其“能够形成何种可计量价值”。评价对象由数据质量逐步扩展至任务效能及经济价值，使数据盒的价值能够在实际使用过程中持续积累和验证。

在企业数据管理中，上述表征结果还可为数据资产台账、成本归集、使用记录及收益分析提供统一的数据对象。数据盒在生产过程中形成的数据来源、治理轨迹、版本信息、质量记录及调用日志，可为数据资源入表、数据资产评价及高质量数据集建设提供相应的过程证据。例如，可按照数据盒归集治理、标注、语义建模及封装成本，并通过调用日志记录数据的使用情况及任务效果；数据盒随生产过程形成的来源、质量、版本及应用说明，也可为后续质量评价和相关认定提供可追溯的信息基础。

因此，面向智能体应用的数据盒不仅承担数据组织和调用功能，还能够作为连接数据治理成果、应用效能及价值评价的统一计量单元。通过将治理过程、调用行为与价值证据关联到同一数据盒，可为数据资产管理、高质量数据集建设及后续数据产品化提供结构化支撑。

4 数据盒的智能化构建方法

当前高质量数据集建设仍较多采用项目制方式推进，不同项目在业务分析、数据接入、治理流程、质量控制及交付方式等方面存在差异，实施过程较依赖项目团队经验，已有成果的复用能力也相对有限。与此同时，传统数据治理主要面向已有数据资源的清洗、标准化及管理，对业务问题形成、需求澄清、规则讨论及方案调整等过程性信息利用不足，而这些信息反映了业务人员及领域专家对业务问题的持续认知与判断，是刻画业务语义及其演化过程的重要依据。

为此，本文引入业务认知轨迹，用于持续记录业务问题提出、关系识别、规则形成及判断修正等认知演化过程，为业务模型构建、数据治理范围确定及后续动态更新提供依据。在此基础上，借鉴工业母机的组织思想提出数据母机（Data

Machine），通过标准化流程协同组织业务建模与数据治理，将业务认知轨迹、业务模型及多源异构数据逐步转化为面向智能体调用的数据盒。数据母机由此构成数据盒智能化构建的核心机制，推动高质量数据建设由一次性项目实施向可重复、可持续的生产过程转变。

4.1 基于业务认知轨迹的双轨构建机制

数据母机是以应用需求为牵引、业务认知轨迹贯穿的数据持续生产机制，整个过程中业务轨与数据轨并行推进、持续交互。

业务轨将一次性业务调研拓展为业务认知过程的持续记录与建模，形成可追溯、可迭代的业务模型；数据轨则依据业务模型确定数据治理对象、范围及粒度，并对多源异构数据开展治理，形成与具体业务任务相适配的高质量数据内容。两条轨道通过业务对象与数据对象之间的映射关系形成双向约束。当业务对象缺乏相应数据支撑时，可据此识别数据缺失或业务模型定义偏差；当数据治理过程中发现新的稳定关系或业务规则时，则可反向补充业务认知轨迹并修正业务模型。通过持续交互，数据治理范围由广泛的数据汇聚逐步收敛为面向特定场景及任务需求的有效数据集合。

该机制具有三个主要特征：一是通过统一业务建模、数据治理及反馈更新过程，提高数据盒构建流程的标准化和可复用性；二是通过持续积累业务认知轨迹，为业务模型更新及智能体后续学习提供过程性数据基础；三是通过业务模型对治理对象及范围进行约束，使数据治理具有明确的场景适配性，并能够随业务变化动态调整。

4.2 数据盒构建流程与更新机制

在双轨协同机制基础上，本文构建面向智能体应用的数据盒智能化构建框架，即数据母机。该框架在数据治理基本流程中引入场景需求约束与运行反馈机制，以具体任务目标确定数据对象、业务规则及治理范围，并将智能体运行过程中产生的查询记录、任务结果、异常信息及用户修正等反馈重新纳入业务认知轨迹，用于持续更新业务模型、治理规则及数据盒配置，从而形成数据盒构建、运行与更新的持续迭代过程。

数据母机包括五个环节：场景界定与接入准备、数据识别与轨迹挖掘、双轨对齐与协同治理、融合封装与服务配置、运行反馈与持续更新。其中，场景界定提供任务约束与初始输入，双轨对齐承担构建阶段的语义校验与治理纠偏，运行反馈则实现运行阶段的数据盒持续更新。数据质量及应用效果评价可结合高质量数据集在说明文档、数据质量和模型应用等方面的相关要求进行设计。

（1）场景界定与接入准备

数据盒构建以具体应用场景和智能体任务为起点。通过业务现状及应用需求调研，明确建设目标、应用方向、业务对象、任务边界、参与角色及预期输出，并收集业务文档、会议纪要、专家意见、调研材料及需求变更记录等业务资料，对场景形成过程中的关键决策进行记录。

与此同时，对与场景相关的业务系统、ERP、CRM、数据库及文档库等数据资源进行梳理，明确数据来源、责任主体、接入范围、接入方式、数据口径及权限要求，形成场景说明、业务材料及数据源接入清单，为后续业务认知轨与数据治理轨的解析提供基础输入。后续环节若发现业务资料或数据资源不足，可返回本环节补充调

研、扩展数据源或调整接入口径。

（2）数据识别与轨迹挖掘

该环节分别对业务认知轨和数据治理轨进行初始解析。业务认知轨对业务文档、会议纪要、调研材料及专家意见等进行结构化处理，提取业务对象、对象关系、业务规则、约束条件及关键决策，形成可追溯的业务认知轨迹和初始业务模型。

数据治理轨依据数据源接入清单，对数据类型、格式、结构、来源、完整性及质量状态进行识别，形成数据资源画像，并据此初步确定数据治理边界、治理粒度及信息补充需求。

当后续校验发现业务模型遗漏、数据缺失或语义映射异常时，本环节负责定位问题来源，并确定业务模型修订、数据补充、治理规则调整或新增数据接入等处理方式；涉及新增业务材料或扩展数据源时，则进一步回溯至场景界定与接入准备环节。

（3）双轨对齐与协同治理

在初始业务模型和数据资源画像基础上，建立业务对象、关系、规则及约束与数据表、字段、文档及数据项之间的语义映射。业务认知轨提供任务语义和场景规则，数据治理轨据此开展数据清洗、标准化、实体匹配、语义对齐、质量治理及组织加工；数据资源状态及治理结果则反向用于校验和修正业务模型。

该环节重点对业务模型、数据资源和任务需求之间的完整性、一致性及适用性进行校验。当业务对象、规则或约束存在建模缺失时，返回上一环节补充业务认知轨迹和业务模型；当存在数据缺失、质量不足或数据内容与业务语义不匹配时，则重新检查数据资源画像、治理范围及数据接入边界。通过持续对齐和校验，使治理范围逐步收敛至能够支撑具体任务的数据集合。

经校验满足任务需求的数据内容、业务语义及治理规则进入融合封装环节。

（4）融合封装与服务配置

在业务模型、数据资源及治理规则达到预设完整性和一致性要求后，将治理后的数据内容、业务语义、自主程序及权限约束按照第3节提出的面向智能体四元数据组织结构进行融合封装，形成数据盒。

封装过程中保留业务模型、业务认知轨迹、源数据及其映射关系，并配置面向智能体调用的服务接口、访问权限、版本信息及质量说明。其中，完整性要求用于保证业务对象、关键规则、必要数据项及其关联关系得到有效覆盖，一致性要求用于保证业务模型、数据内容、治理规则及权限配置之间保持协调。

通过该环节，数据盒形成面向具体任务的数据供给、业务语义表达、受控调用及版本追溯能力。

（5）运行反馈与更新

数据盒投入运行后，持续采集智能体调用过程中产生的查询记录、任务结果、异常信息、用户修正及业务反馈，并根据反馈类型触发相应更新机制。对于任务目标、业务规则或数据范围发生的较大变化，反馈信息回流至场景界定环节；对于业务模型、数据内容、语义映射及治理规则的调整需求，回流至数据识别及双轨协同治理环节；对于接口、权限及版本等配置变化，则进入融合封装与服务配置环节进行更新。

构建阶段的校验回溯与运行阶段的反馈更新共同形成数据盒的持续迭代机制，使业务认知轨迹、业务模型、治理规则及数据内容能够随场景和任务变化动态更新，并支持数据盒的持续演化与复用。

在部署方式上，数据母机可运行于组织内部或专有环境，使原始数据在受控边

界内完成识别、治理及封装。智能体通过数据盒提供的标准化接口访问数据服务，降低对底层原始数据的直接访问，从而兼顾数据调用能力与数据安全要求。

5 总结与展望

本文面向智能体在垂直领域的规模化应用，针对现有数据治理与具体业务任务衔接不足的问题，从数据组织与数据构建两个层面提出面向智能体的数据盒重构及智能化构建方法。在数据组织层面，通过引入业务语义、业务规则及任务上下文，形成面向智能体调用的数据盒结构；在数据构建层面，通过业务认知轨迹及双轨协同机制，实现业务模型、数据治理与数据盒构建的持续关联。所提方法为智能体提供具有任务适配性、业务语义及可调用性的数据组织基础，并可与企业既有数据基础设施衔接，为高质量数据集构建及数据资产管理提供统一的工程路径。

后续研究将进一步围绕方法验证与规模化应用展开。一方面，在典型行业场景中验证数据盒构建效率、任务适配效果、更新能力及跨场景复用能力；另一方面，研究业务认知轨迹、业务模型、治理规则及数据映射关系的行业级积累与复用机制，形成由组织级专识数据集与行业通识数据集协同构成的数据供给体系，为智能体应用提供持续、可复用的数据支撑。

参考文献：

[1] Gadre S, Ilharco G, Fang A, et al. DataComp: In search of the next generation of multimodal datasets[C]//Advances in Neural Information Processing Systems 36. 2023: 27092-27112.

[2] Li J, Fang A, Smyrnis G, et al. DataComp-LM: In search of the next generation of training sets for language models[C]//Advances in Neural Information Processing Systems 37. 2024: 14200-14282.

[3] Shumailov I, Shumaylov Z, Zhao Y, et al. AI models collapse when trained on recursively generated data[J]. Nature, 2024, 631: 755-759.

[4] 李继峰,张成龙,刘鑫;等.面向人工智能的数据治理框架[J].大数据,2025,11(01):3-20. DOI：10.11959/j.issn.2096-0271.2025004.
LI Jifeng, ZHANG Chenglong, LIU Xin, et al. Data governance framework for artificial intelligence[J]. BIG DATA RESEARCH, 2025, 11(01): 3-20. DOI：10.11959/j.issn.2096-0271.2025004.

[5] 陈璐, 郭宇翔, 葛丛丛, 郑白桦, 高云君. 基于联邦学习的跨源数据错误检测方法[J]. 软件学报, 2023, 34(3): 1126-1147.
CHEN Lu, GUO Yu-Xiang, GE Cong-Cong, ZHENG Bai-Hua, GAO Yun-Jun. Cross-source Data Error Detection Approach Based on Federated Learning[J]. Journal of Software, 2023, 34(3): 1126-1147.

[6] Jarrahi M H, Memariani A, Guha S. The principles of data-centric AI[J]. Communications of the ACM, 2023, 66(8): 84-92.

[7] Yan M, Wang Y, Pang K, et al. Efficient Mixture of Experts based on Large Language Models for Low-Resource Data Preprocessing[C]//Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: ACM, 2024: 3690-3701.

[8] 任磊, 董家宝, 曾宪超, 等. 数字族谱: 驱动工业具身智能世界模型. 中国科学: 信息科学, 2025, 55: 1748 - 1765, doi: 10. 1360/SSI-

2025-0093

Ren L, Dong J B, Zeng X C, et al. Digital genealogy: empowering industrial embodied intelligence world model. Sci Sin Inform, 2025, 55: 1748 - 1765, doi: 10.1360/SSI-2025-0093

[9] Anokhin P, Semenov N, Sorokin A, et al. AriGraph: Learning Knowledge Graph World Models with Episodic Memory for LLM Agents[C]//Proceedings of the 34th International Joint Conference on Artificial Intelligence. 2025: 12-20.

[10] Guan L, Valmeekam K, Sreedharan S, et al. Leveraging Pre-trained Large Language Models to Construct and Utilize World Models for Model-based Task Planning[C]//Advances in Neural Information Processing Systems 36. 2023: 79081-79094.

[11] 熊贇, 朱扬勇. 面向数据自治开放的数据盒模型[J]. 大数据, 2018,4(2):2018015. DOI：10.11959/j.issn.2096-0271.2018015.

Yun XIONG, Yangyong ZHU. Data box:a novel data model for self-governing openness of data[J]. Big data research, 2018, 4(2): 2018015. DOI：10.11959/j.issn.2096-0271.2018015.

[12] 叶雅珍, 朱扬勇. 盒装数据:一种基于数据盒的数据产品形态[J]. 大数据, 2022,8(3):15-25. DOI：10.11959/j.issn.2096-0271.2022030.

Yazhen YE, Yangyong ZHU. BoxedData: a data product form based on databox[J]. Big data research, 2022, 8(3): 15-25. DOI：10.11959/j.issn.2096-0271.2022030.

[13] 熊贇,朱扬勇.数据产品及其流通监管体系研究[J].大数据,2025,11(03):98-107. DOI：10.11959/j.issn.2096-0271.2025042.

XIONG Yun, ZHU Yangyong. Research on data product and their circulation regulatory framework[J]. BIG DATA RESEARCH, 2025, 11(03): 98-107. DOI：10.11959/j.issn.2096-0271.2025042.

[14] 吴俊,谢文静,熊贇.数据产品的统一标识编码方法研究[J].大数据,2026,12(02):54-63. DOI：10.11959/j.issn.2096-0271.2026035.

Wu Jun,Xie Wenjing,Xiong Yun.Research on unified identification for data products[J]. BIG DATA RESEARCH, 2026, 12(02): 54-63. DOI：10.11959/j.issn.2096-0271.2026035.

[15] 杨琳,朱扬勇.高质量数据集产品的形态和生产流程研究[J].大数据,2026,12(02):64-74. DOI：10.11959/j.issn.2096-0271.2026036.

Yang Lin,Zhu Yangyong.Research on the form and production process of high-quality dataset products[J]. BIG DATA RESEARCH,2026,12(02):64-74. DOI：10.11959/j.issn.2096-0271.2026036.

[16] Ouyang S, Huang J, Pillai P, et al. Ontology Enrichment for Effective Fine-grained Entity Typing[C]//Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: ACM, 2024: 2318-2327.

作者简介



谷婧雯（1987– ），女，硕士，上海所迹数字科技有限公司，中级经济师，总经理，主要研究方向为数据科学和数字经济。



熊赟（1980– ），女，博士，复旦大学计算与智能创新学院教授，上海市数据科学重点实验室副主任，主要研究方向为数据科学和数字经济。



朱扬勇（1963– ），复旦大学教授，上海数据研究院有限公司学术副院长，《大数据》期刊编委会副主任。数据科学家，在数据领域做出开创性工作，提出数据自治、数据财政、数据资产等概念和体系。在数据领域发表论文200多篇，持有高价值专利15件，出版学术著作17本。上海市数据科学重点实验室创始主任，担任多个政府顾问。主导编制了多个地方政府数据发展规划。主要研究方向为数据科学与数字经济。

录用日期: XXXX-XX-XX
通信作者:
基金项目:
Foundation Items: