Data Agents: Levels, State of the Art, and Open Problems

Yuyu Luo[†]
HKUST (GZ)
Guangzhou, China
yuyuluo@hkust-gz.edu.cn

Guoliang Li[†]
Tsinghua University
Beijing, China
liguoliang@tsinghua.edu.cn

Ju Fan
Renmin University of China
Beijing, China
fanj@ruc.edu.cn

Nan Tang
HKUST (GZ)
Guangzhou, China
nantang@hkust-gz.edu.cn

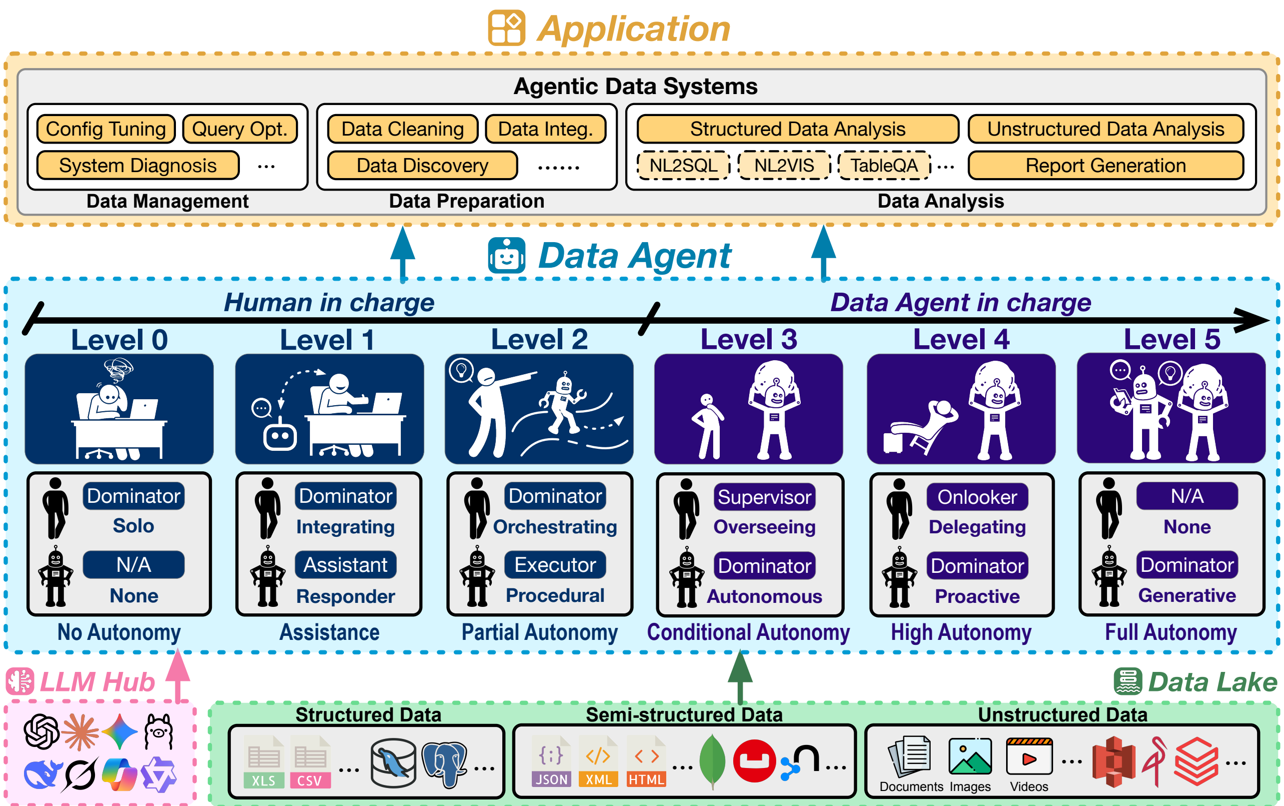

Figure 1: An Overview of Data Agents. (https://github.com/HKUSTDial/awesome-data-agents)

Abstract

Data agents leverage large language models (LLMs) and tool-using agents to automate data management, preparation, and analysis. However, the term is currently used inconsistently, blurring the boundaries of capability and accountability.

In this tutorial, we propose the first hierarchical taxonomy of data agents from Level 0 (L0, no autonomy) to Level 5 (L5, full autonomy). Building on this taxonomy, we will introduce a lifecycle- and level-driven view of data agents. We will (1) present the L0–L5 taxonomy and the key evolutionary leaps that separate simple assistants from truly autonomous data agents, (2) review representative L0–L2 systems across data management, preparation, and analysis, (3) highlight emerging Proto-L3 systems that strive to autonomously orchestrate end-to-end data workflows to tackle diverse and comprehensive data-related tasks under supervision, and (4) discuss forward-looking research challenges towards proactive (L4) and generative (L5) data agents. We aim to offer both a practical map of today's systems and a research roadmap for the next decade of data-agent development.

CCS Concepts

• **Information systems → Data management systems**; • **Computing methodologies → Artificial intelligence**.

Keywords

Data Agents; Large language models; LLM Agents

ACM Reference Format:
Yuyu Luo[†], Guoliang Li[†], Ju Fan, and Nan Tang. 2026. Data Agents: Levels, State of the Art, and Open Problems. In *Companion of the International Conference on Management of Data (SIGMOD Companion '26), May 31-June 05, 2026, Bengaluru, India.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3788853.3801878

This work is licensed under a Creative Commons Attribution 4.0 International License.
SIGMOD Companion '26, Bengaluru, India
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2450-3/2026/05
https://doi.org/10.1145/3788853.3801878

1 Introduction

Modern data ecosystems are increasingly complex, spanning heterogeneous and multimodal data sources, evolving schemas, and tightly coupled Data+AI pipelines [35, 36, 41, 104]. At the same time, LLM-based agents have demonstrated strong capabilities in tool use, planning, and iterative reasoning [42, 49, 52, 76, 95, 106]. As a result, the term *data agents* has rapidly gained popularity in both academia and industry [17, 33, 47, 50, 70], with systems ranging from simple SQL or BI chatbots to ambitious products marketed as fully autonomous "data scientists" [42, 93].

Without a shared vocabulary, however, fundamentally different systems are being conflated under a single, overloaded term. This leads to mismatched user expectations, ambiguous accountability when failures occur, and difficulty in objectively comparing different approaches. Similar challenges were previously faced by the driving-automation community, which motivated the SAE J3016 standard that introduced a six-level taxonomy of autonomy [62].

To address this confusion in data systems community, recent work proposes a hierarchical taxonomy of data agents [108], from Level 0 (L0, no autonomy) to Level 5 (L5, full autonomy), together with a structured survey of existing systems along this axis, which describes how task dominance and responsibility gradually shift from human operators to data agents as autonomy increases.

In this tutorial, we build on that survey and turn it into a *teaching-oriented* framework for SIGMOD attendees. Our goal is to help participants (1) understand what different levels of data agents can realistically do, (2) navigate the growing landscape of systems across the data lifecycle, and (3) identify key research challenges for advancing data agents towards higher autonomy.

1.1 Tutorial Overview

We will give a 3-hour tutorial consisting of a 140-minute lecture-style part (Parts I–IV) followed by a 40-minute *Data Agent Playground* (Part V) for hands-on exploration and discussion.

Part I: Problem Definition and Preliminaries (30 minutes). We begin by motivating data agents in modern Data+AI ecosystems and formalizing the concept of a data agent. We will: (i) introduce the motivation and problem definition of data agents, emphasizing why existing "data assistant" systems are insufficient and why autonomy and responsibility need to be made explicit; (ii) define data agents more formally and contrast them with general-purpose LLM agents; (iii) summarize key challenges (terminology ambiguity, lifecycle fragmentation, autonomy *vs.* governance, technical bottlenecks) and motivate the need for a level-based taxonomy of data agents.

Part II: L0–L2 Data Agents Across the Data Lifecycle (40 minutes). Next, we focus on the lower autonomy levels (L0–L2) and instantiate them in three phases of the data lifecycle: data management, data preparation, and data analysis. We will: (i) give an overview of how L0, L1, and L2 manifest in each phase and connect them to the roles of humans and agents illustrated in Figure 1; (ii) deep-dive into each phase: in data management, from manual DBAs (L0) to database tuning/diagnosis/query optimization copilots (L1) and L2 agents with direct access to DBMSs and monitoring signals; in data preparation, from scripts and rules (L0), to suggestion-style copilots to conduct data cleaning, integration, and discovery (L1), to L2 agents that invoke external tools and close the loop via execution feedback; in data analysis, from structured data analysis (Table QA / NL2SQL / NL2VIS), unstructured data analysis, and report generation with prompt-response paradigm (L1) to L2 environment-perceived analysis agents that maintain state and invoke SQL, plotting, and retrieval tools; (iii) use one or two running examples (e.g., database operations and BI analytics) to make the differences between L0, L1, and L2 concrete. We conclude this part by summarizing recurring design patterns at L0–L2 and their reliability boundaries.

Part III: L3 Data Agents and Proto-L3 Systems (45 minutes). We then move to Level 3, the ongoing research frontier where data agents start to act as workflow orchestrators under human supervision. We will: (i) formally define L3 and explain the key evolutionary leap from L2 to L3; (ii) present representative Proto-L3 systems from academia that explore LLM orchestrators, semantic operators, task DAG optimization, and tool evolution to support versatile, cross-task workflows, and discuss their architectures, supported tasks, orchestration strategies, and limitations; (iii) analyze industrial "data agent" products in cloud data platforms and lakehouses, map them onto corresponding levels, and highlight common design patterns (e.g., DAG-based pipeline orchestration, planner–executor separation, multi-agent collaboration mechanism) and current bottlenecks (e.g., predefined operators/tools, limited causal/meta reasoning, constrained task coverage, strong reliance on human-crafted guardrails).

Part IV: Towards L4–L5 and Research Roadmap (25 minutes). Finally, we complete the lecture part by discussing the visionary Levels 4 and 5 and outlining a research roadmap. We will: (i) elaborate the vision of L4 data agents as proactive, long-lived, self-governing components that continuously monitor Data+AI ecosystems, autonomously discover issues and opportunities, and orchestrate pipelines without explicit instructions; (ii) introduce L5 data agents as generative data scientists that can invent new solutions, algorithms, and paradigms rather than only applying existing methods; (iii) summarize key open problems, including autonomous orchestration and versatility, causal and meta reasoning, intrinsic motivation and task discovery, long-horizon planning and trade-offs, safety and governance, and benchmarks for autonomy.

Part V: Data Agent Playground — Hands-on Exploration and Discussion (40 minutes). The final part is an interactive *Data Agent Playground* that increases audience interaction. We will walk through a few concrete data-agent workflows [12, 20, 64, 68, 96], show how L1/L2/Proto-L3 agents behave step by step, and invite attendees to try out our own data-agent prototypes. Participants will be encouraged to sketch or refine agents for their own settings, position them on the L0–L5 spectrum, and discuss key trade-offs in autonomy, governance, and reliability, followed by a brief Q&A that ties these insights back to the research roadmap in Part IV.

1.2 Our Scope and Goals

Our Distinction from Existing Tutorials. Existing tutorials and surveys on LLMs and data systems typically focus on specific aspects such as LLMs for databases and data analysis [35, 36, 43, 46, 75,

†Yuyu Luo and Guoliang Li are the corresponding authors.

Table 1: Comparison between General LLM Agents and Data Agents

Aspect	General LLM Agents	Data Agents
Primary Focus	Task and Content Centric: *Completing defined tasks or generating content.*	Data-Lifecycle Centric: *Data management, preparation, and analysis.*
Problem Scope	Self-contained and Static: *Acts on explicit instructions and a finite prompt.*	Exploratory and Dynamic: *Actively explores and navigates vast, dynamic data lakes.*
Input Data	Small-Scale and Ready-to-Use: *Typically receives manageable, clean inputs.*	Large-Scale and "Raw": *Designed to handle heterogeneous, dynamic, and noisy raw data.*
Tool Invocation	General-Purpose Toolkit: *Web search, calculators, OCR, image generators, etc.*	Specialized Data Toolkit: *DB loaders, SQL equivalence checker, visualization libraries, etc.*
Primary Output	Generative Artifacts: *Human-consumable product: dialogues, reasoning, images, etc.*	Data Products and Insights: *Config, processed data, insights, visualizations, analytical reports, etc.*
Error Consequence	Localized: *Typically affects limited to only the direct output.*	Cascading: *Errors can cascade, affecting downstream insights.*

102, 103], data management for machine learning [8, 15, 92, 104], or general-purpose LLM agents and tool-using systems [68]. In contrast, our tutorial is distinguished by three aspects:

(1) **Level-based view.** We adopt a *level-based* perspective on data agents (L0–L5) that explicitly links autonomy, capability, and responsibility, making it easier to reason about what a "data agent" at each level can and cannot do.
(2) **Holistic lifecycle perspective.** We take a *holistic lifecycle* view, jointly covering data management, data preparation, and data analysis under a unified data-agent framework, rather than treating individual tasks in isolation.
(3) **Evolutionary leaps and roadmap.** We emphasize the *evolutionary leaps* between levels, especially the crucial L2→L3 and L3→L4 transitions, and present a *research roadmap* towards proactive (L4) and generative (L5) data agents, instead of providing an exhaustive but flat catalogue of systems.

Target Audience and Learning Outcomes. This tutorial is intended for a broad SIGMOD audience, including researchers in databases, data mining, machine learning, AI agents, and data-centric AI; system developers and practitioners building data platforms, lakehouses, or enterprise data stacks; and students who wish to enter the emerging area of data agents. By the end of the tutorial, participants will be able to use the L0–L5 framework to position existing and future systems, distinguish data agents from general-purpose LLM agents, interpret and calibrate vendor claims about "data agents", choose appropriate autonomy levels for their own applications, and reason about key design dimensions such as perception, planning, tools, memory, and governance. We assume familiarity with basic database concepts and LLM terminology; the tutorial itself will be self-contained.

2 Tutorial Outline

2.1 Background and Problem Definition

2.1.1 Problem Description: What is a Data Agent? Informally, a *data agent* is an LLM-based architecture that orchestrates a Data+AI ecosystem to perform data-related tasks such as configuration tuning, data cleaning, integration, exploration, and analysis [17, 70, 103]. Formally, we can define a data agent $\mathcal{A}$ that operates on raw data $\mathcal{D}$ within an environment $\mathcal{E}$ (*e.g.,* DBMS, code interpreters, APIs, etc.), utilizing LLMs $\mathcal{M}$, ultimately producing an output O to tackle the data-related task $\mathcal{T}$, abstractly represented as: $\mathcal{A} : (\mathcal{T}, \mathcal{D}, \mathcal{E}, \mathcal{M}) \rightarrow O$.

2.1.2 Task Landscape and Data Agents vs. General LLM Agents. Data agents operate within modern Data+AI ecosystems that span relational databases, data warehouses and lakehouses, data lakes, ETL/ELT pipelines, BI tools, and ML services. Therefore, data agents must reason over large, heterogeneous, and often schema-rich data lakes without exhaustive ingestion [36]; interact with dynamic and noisy data and systems whose behavior changes over time [35]; and operate inside multi-stage pipelines where errors can silently propagate and amplify, rather than affecting only a single response.

Compared to general-purpose LLM agents, data agents thus face more constrained yet substantially more demanding environments. They also need to satisfy stringent requirements on reliability, governance, and reproducibility that are less prominent in many generic agent settings. Table 1 summarizes key differences between data agents and general LLM agents along these dimensions.

2.2 The L0–L5 Hierarchy of Data Agents

Inspired by the SAE J3016 standard for driving automation [62], we adopt a six-level taxonomy of data agents from L0 to L5. As summarized in Figure 1, data agents are organized into six autonomy levels, from Level 0 (L0) to Level 5 (L5). The figure indicates, for each level, who is in charge of the data-related task (human *vs.* data agent), what role the data agent plays (*e.g.,* responder, executor, orchestrator, proactive or generative component), and which parts of the data lifecycle (management, preparation, analysis) are involved.

As an overview, Figure 2 positions representative systems from academia and industry across the L0–L5 levels and the three phases of the data lifecycle. We briefly review these levels below.

L0: No Autonomy. At L0, there is no data agent involvement. All tasks in data management, preparation, and analysis are performed manually by humans.

L1: Assistance. L1 data agents operate within a stateless, prompt-response framework. They can answer questions, generate code snippets, or suggest queries, but they do not perceive or interact with the environment. Humans remain fully responsible for executing and verifying any suggestions.

L2: Partial Autonomy. L2 data agents gain the ability to perceive and interact with their environment, including data lakes, DBMSs, code interpreters, and external APIs. They may possess memory and can invoke tools to autonomously execute task-specific procedures within human-orchestrated pipelines.

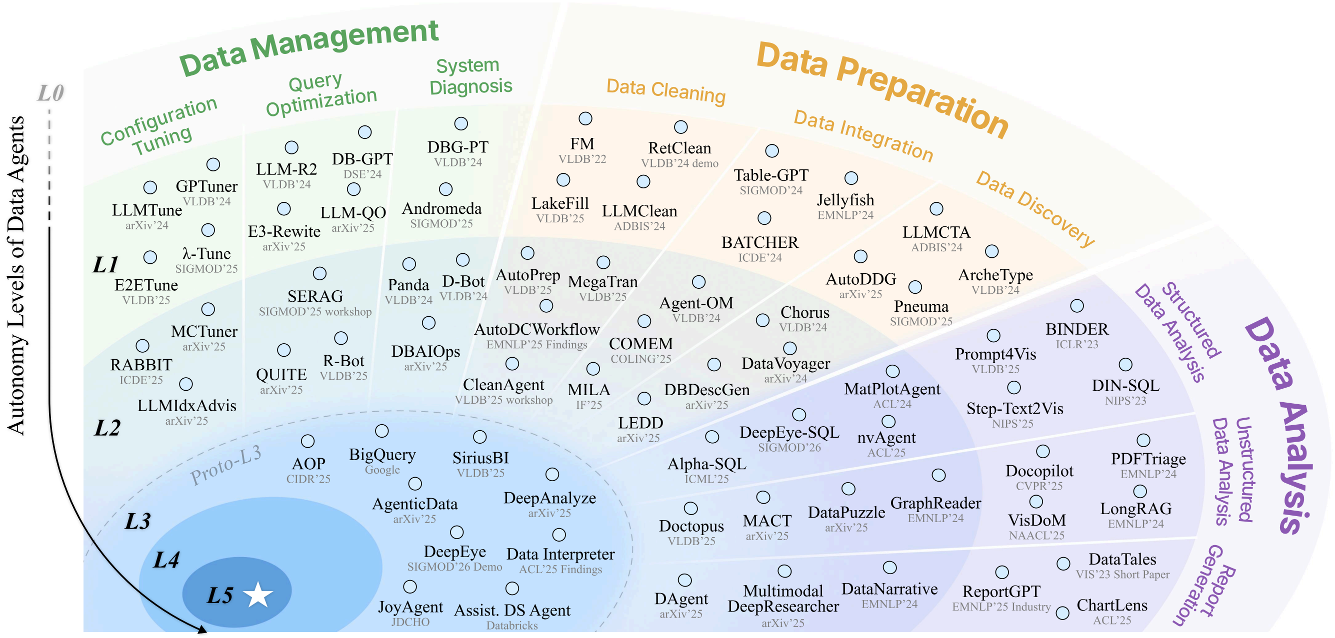

Figure 2: Representative Data Agents Across Different Levels.

L3: Conditional Autonomy. L3 data agents are expected to autonomously orchestrate and execute tailored data pipelines for a wide range of tasks under human supervision. They interpret high-level user intentions and dominate the end-to-end workflow, while humans act as supervisors.

L4: High Autonomy. L4 data agents achieve high autonomy and reliability, eliminating the need for human supervision and explicit instructions. They are fully delegated to proactively monitor Data+AI ecosystems, autonomously discover issues and opportunities in data lakes, and orchestrate pipelines to address them.

L5: Full Autonomy. At L5, data agents are envisioned to innovate new solutions and paradigms beyond existing methods, acting as fully autonomous and generative data scientists. Human involvement becomes unnecessary.

2.3 L0–L2: From Manual Workflows to Partial Autonomy

In this section, we review representative systems at L0–L2 across three phases of the data lifecycle: data management, data preparation, and data analysis.

2.3.1 Data Management. Data management includes configuration tuning, query optimization, and system diagnosis in database systems [100, 104]. At **L0**, DBAs manually tune knobs, index configurations, and execution plans, relying on expertise and trial-and-error [100]. At **L1**, LLMs are used as query-responsive assistants to generate tuning suggestions or rewritten queries. They operate in a prompt-response manner, returning recommendations that humans must integrate and validate [18, 28, 40]. For instance, λ-Tune [19] and E2ETune [23] use LLMs to recommend configuration candidates based on workload features, and Andromeda [9] generates diagnostic suggestions for configuration debugging. At **L2**, data agents gain direct access to the DBMS and monitoring information. They can observe workload statistics, execute tuning experiments, and adjust configurations or rewrite queries in a decision loop, while still following human-designed workflows [65, 86, 101]. Rabbit [69], R-Bot [71], D-Bot [105] exemplify this through utilizing environmental feedback in configuration tuning, query rewriting, and system diagnosis.

2.3.2 Data Preparation. Data preparation [7, 15, 34] covers data cleaning [4], integration [37], and discovery [16]. At **L1**, data agents primarily act as suggestion engines: RetClean [53] and LakeFill [87] infer and impute missing values, LLMClean [4] generates rules for cleaning tasks, Narayan et al. [54] deploy LLMs to propose schema matches or entity correspondences, AutoDDG [94] and LLMCTA [27] produce dataset summaries, metadata, or column annotations. Homomorphic compression [21] is a promising method for reducing the computational cost of data agents while maintaining semantic integrity. At **L2**, data agents go beyond query responder and directly interact with databases or data lakes to execute cleaning and transformation operations, verify constraints, and adjust their strategies based on execution feedback, and iteratively refine integration decisions as more data is explored [26, 59, 81]. Representative systems include CleanAgent [59], MegaTran [34] for data cleaning; SEED [10], Agent-OM [60] for data integration; LEDD [3] and DBDescGen [39] for data discovery.

2.3.3 Data Analysis. Data analysis includes structured and unstructured data analysis, as well as report generation. At **L1**, we mostly see LLM-driven question-answering assistants for Table QA [11, 66, 90], NL2SQL [30, 44, 58, 97, 98, 107], NL2VIS [38, 45, 48, 51, 84], textual or multimodal Document QA [14, 61, 72, 82, 83], which generate answers to respond to user questions over curated datasets, and report generators [6, 67, 73] that operate on input tables or documents. At **L2**, data agents move beyond static querying to dynamically engage with, verify, and refine multi-step analytical processes [57, 63, 77, 91]. They invoke tools such as SQL

Table 2: Comparison of Representative Proto-L3 Data Agents from Academia Research and Industry Products. Compares Open-source availability; Undef Ops.: capabilities in utilizing non-predefined operators; data-related task coverage across data management, preparation, analysis; data complexity dimensions: Multi-source (Multis.), Heterogeneous (Hete.), and Multimodal (Multim.)

Years	Data Agent	Open-source	Undef Ops.	Data Complexity			Data Management			Data Preparation			Data Analysis		
				Multis.	Hete.	Multim.	Config Tun.	Query Opt.	Sys. Diag.	Data Clean.	Data Integ.	Data Disc.	Struct.	Unstruct.	Report Gen.
2026	DeepEye [31]	✓	✗	✓	✓	✓	-	-	-	✓	✓	✓	✓	✓	✓
2025	AgenticData [68]	-	✗	✓	✓	-	-	✓	✓	✓	✓	✓	✓	✓	-
2025	DeepAnalyze [96]	✓	-	✓	✓	-	-	-	-	✓	✓	✓	✓	✓	✓
2025	AOP [78]	-	-	✓	✓	✓	-	✓	✓	✓	✓	✓	✓	✓	-
2025	iDataLake [79]	✓	-	✓	✓	✓	-	✓	-	✓	✓	✓	✓	✓	✓
2024	Data Interpreter [22]	✓	-	-	✓	✓	-	-	-	✓	-	✓	✓	✓	✓
2025	JoyAgent [2]	✗	✗	✓	✓	-	-	-	✓	✓	✓	✓	✓	✓	✓
2025	Assist. DS Agent [12]	-	-	✓	✓	-	-	✓	-	✓	✓	✓	✓	✓	✓
2025	TabTab [74]	-	-	✓	✓	-	-	-	-	✓	✓	-	✓	✓	✓
2025	ByteDance Data Agent [5]	-	-	✓	✓	-	-	-	-	-	✓	-	✓	✓	✓
2025	BigQuery [20]	-	-	✓	✓	-	-	✓	-	✓	✓	✓	✓	-	-
2025	Cortex [64]	-	-	✓	✓	✓	-	-	-	✓	✓	✓	✓	✓	-
2025	Xata Agent [1]	-	-	✓	✓	-	✓	✓	✓	-	-	✓	-	-	-
2025	SiriusBI [24]	-	-	-	-	-	-	-	-	✓	-	✓	✓	-	✓

engines [29, 32], plotting libraries [55, 85, 89], or retrieval modules [25, 80, 99], and support iterative exploration and refinement of analyses [13, 56, 88]. For example, DeepEye-SQL [29] improves Text-to-SQL reliability through a software-engineering-style workflow with schema grounding, multi-path SQL generation, deterministic checker-based verification, and confidence-aware selection.

2.4 L3: Striving for Autonomous Data Agents

We now turn to Level 3 (L3), which marks a crucial step from procedural executors to autonomous orchestrators.

2.4.1 From Executor to Dominator. At L2, humans design the overall pipelines, and data agents execute specific procedures within these human-prescribed workflows. At L3, by contrast, data agents are expected to interpret high-level user intent and autonomously orchestrate pipelines that span data management, preparation, and analysis. During execution, data agents adapt the pipeline based on feedback and intermediate results, while humans primarily act as supervisors who review plans and outcomes rather than as pipeline designers. In this sense, task dominance and primary responsibility shift from humans to data agents. Figure 3 illustrates the typical L3 data agent, highlighting its conditional autonomy in autonomous pipeline orchestration and optimization.

2.4.2 Proto-L3 Data Agents in Research. Recent research systems begin to exhibit partial L3 capabilities. They use LLM-based orchestrators [68], predefined operators [78, 79], workflow optimization [22, 78], and tool libraries [96] to orchestrate multi-step workflows over heterogeneous systems, cover multiple stages of the data lifecycle within a single agentic process, and maintain state across long-running interactions so that they can refine their plans and correct mistakes over time. DeepEye [31][1] exemplifies this direction by orchestrating data analysis as a transparent DAG over heterogeneous sources such as databases, files, and knowledge bases, combining multimodal orchestration, hierarchical reasoning, and workflow-level validation and optimization under human supervision. These Proto-L3 agents typically operate in constrained

[1]DeepEye: https://deepeye.hk/

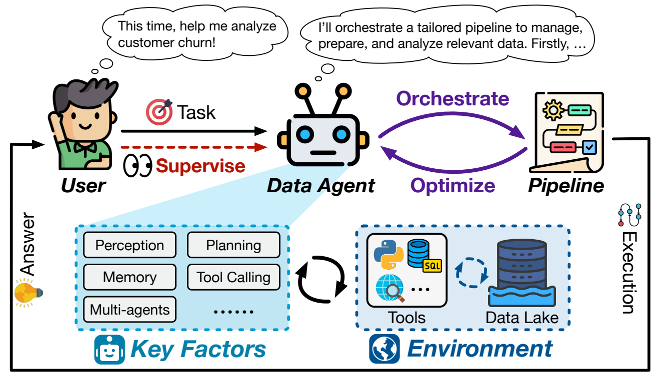

Figure 3: L3 Data Agents (Conditional Autonomy).

environments with curated tools and data, but they provide concrete testbeds for studying the transition from execution-focused L2 agents to orchestration-centered L3 systems.

We will present several representative academic systems and discuss: (i) their pipeline representation, orchestration, and optimization strategies; (ii) their architectural choices (single vs. multi-agent, central vs. decentralized planners); (iii) their approach to tool abstraction and composition; and (iv) their strategies for incorporating feedback and handling errors.

Table 2 compares representative Proto-L3 data agents from both academia and industrial products along dimensions such as tool flexibility, data complexity, data lifecycle coverage, and specific management, preparation, and analysis tasks they support.

2.4.3 Industrial Data-Agent Products. Industrial platforms (*e.g.,* cloud data warehouses and lakehouses) have started to offer commercial "data agent" products [20, 64]. We analyze: (i) how these products map to the L0–L3 levels in practice; (ii) which guarantees they provide (*e.g.,* human-in-the-loop confirmation, logging, and rollback); and (iii) common limitations and design trade-offs.

2.4.4 Current Bottlenecks and Gaps. This tutorial identifies several gaps preventing full L3 autonomy: (1) limited pipeline orchestration capabilities and reliance on predefined operators; (2) inadequate higher-order, causal, and meta-reasoning to diagnose cascading errors; (3) difficulty adapting to dynamic environments with changing data and workloads; and (4) heavy reliance on human-crafted reinforcement learning setups for alignment. These challenges motivate new methods beyond straightforward tool-calling LLM agents.

2.5 L4–L5: Vision and Research Roadmap

Finally, we discuss the Levels 4 and 5 and outline a research roadmap.

2.5.1 L4: Proactive, High-Autonomy Data Agents. At L4, data agents are envisioned as proactive, long-lived, and self-governing components of Data+AI ecosystems. Instead of merely reacting to explicit user requests, an L4 agent continuously monitors data lakes, systems, and models, detects phenomena such as data drift, performance regressions, and schema changes, and identifies opportunities such as beneficial materializations, missing indexes, or promising analytical workflows. It is expected to prioritize among these tasks, design and adapt pipelines to address them without explicit instructions, and operate within reliability, safety, and governance constraints even in the absence of human supervision. Typical scenarios include autonomous detection and mitigation of workload shifts, long-horizon management of indexes and materialized views, and continuous quality assurance for critical data assets. Realizing such capabilities not only raises questions about autonomous orchestration across the full data lifecycle but also calls for mechanisms for intrinsic motivation, task discovery in large data ecosystems, and long-horizon planning that reasons about cumulative cost, latency, and data-quality trade-offs.

2.5.2 L5: Generative Data Agents. L5 data agents go beyond deploying existing techniques and are conceived as autonomous, generative data scientists. An L5 data agent is expected to identify gaps in current methods, hypothesize new algorithms or representations when existing approaches are insufficient, design and analyze experiments to test these hypotheses, and iteratively refine its own solutions over time. In this vision, the data agent is not only a user of database and ML systems, but also an active contributor to their evolution. Moving towards L5 requires abstractions that allow data agents to manipulate high-level design choices—such as physical designs, query rewrite strategies, data cleaning policies, or learning procedures—while staying grounded in executable systems, as well as causal and meta reasoning supporting principled diagnosis, comparison, and improvement of alternative designs, even pioneering of innovative solutions, novel theories, and new paradigms.

Although fully realized L4 and L5 data agents remain speculative, articulating these levels helps delineate a research roadmap. In the near term, the most pressing challenges lie in making L2 and Proto-L3 agents more robust, transparent, and governable; in the medium term, progress toward L4 will depend on advances in autonomous orchestration, task discovery, and long-horizon decision making under multi-objective constraints; and in the longer term, movement toward L5 will hinge on integrating causal and meta reasoning with agent-driven experimentation and on developing evaluation methodologies that capture autonomy, adaptability, and safety beyond traditional task-level accuracy.

2.5.3 Research Opportunities. The L0–L5 hierarchy suggests several research directions that are closely tied to core data management problems. A central question is how data agents should perceive and act over large, heterogeneous data lakes: which indexes, materialized views, summaries, or learned representations should serve as their "senses", how these structures are exposed as tools, and how agents can orchestrate complex pipelines across management, preparation, and analysis while preserving performance, data quality, and governance guarantees.

A second theme concerns how data agents are trained and evaluated in realistic environments. Here, operational logs, configuration histories, and telemetry can form the basis for constructing training corpora, adapting agent policies over time, and supporting causal and meta reasoning about failures and improvements. This, in turn, calls for benchmarks and methodologies that go beyond task-level accuracy to capture autonomy, robustness, adaptability, and safety on realistic data-management workloads.

3 BIOGRAPHY

Yuyu Luo is an Assistant Professor at The Hong Kong University of Science and Technology (Guangzhou), with an affiliated position at the HKUST. His research interests include Data Agents, AI4DB, and Data-centric AI. He has received the Best-of-SIGMOD 2023 Papers. He organized the first Agentic Data Systems workshop at VLDB 2026 and the first Data Agent Competition at KDD Cup 2026.

Guoliang Li is a full professor in the Department of Computer Science, Tsinghua University. His research interests mainly include data cleaning and integration, and machine learning for databases. He got the VLDB 2017 early research contribution award, TCDE 2014 Early Career Award, VLDB 2023 Industry Best Paper Runner-up, Best of SIGMOD 2023, SIGMOD 2023 research highlight award, DASFAA 2023 Best Paper Award, and CIKM 2017 Best Paper Award.

Ju Fan is a Professor at the DEKE Lab, MOE China, and the School of Information, Renmin University of China. He received his PhD from Tsinghua University in 2013 and received the ACM China Rising Star Award and the 2023 SIGMOD Research Highlight Award. Dr. Fan's main research interests are AI4DB and database systems.

Nan Tang is an Associate Professor at The Hong Kong University of Science and Technology (Guangzhou), with an affiliated position at the HKUST. He has received the VLDB 2010 Best Paper Award, the 2023 SIGMOD Research Highlight Award, and the Best-of-SIGMOD 2023. His main research interests are AI4DB and data-centric AI.

Acknowledgment

This paper was supported by National Key R&D Program of China (2023YFB4503600), NSF of China (62402409, 62525202, 62232009, 62436010, 62441230), Shenzhen Project (CJGJZD20230724093403007), Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM509), China Railway Science Research Institute Group Co., Ltd, Zhongguancun Lab, Huawei, Beijing National Research Center for Information Science and Technology, Scientific Research Innovation Capability Support Project for Young Faculty (Grant No. SRICSPYF-ZY2025001), Guangzhou Basic and Applied Basic Research Foundation (2026A1515010269, 2025A04J3935, 2023A1515110545), and Guangdong Provincial Project (2023CX10X008).

References

[1] [n. d.]. *Xata Agent.*

[2] Agent Team at JDCHO. [n. d.]. *JoyAgent-JDGenie.*

[3] Qi An, Chihua Ying, Yuqing Zhu, Yihao Xu, Manwei Zhang, and Jianmin Wang. 2025. LEDD: large language model-empowered data discovery in data lakes. *arXiv preprint arXiv:2502.15182* (2025).

[4] Fabian Biester, Mohamed Abdelaal, and Daniel Del Gaudio. 2024. LLMClean: Context-Aware Tabular Data Cleaning via LLM-Generated OFDs. arXiv:2404.18681 [cs.DB]

[5] ByteDance Volcengine. [n. d.]. *Data Agent.*

[6] Lucas Cecchi and Petr Babkin. [n. d.]. ReportGPT: Human-in-the-loop Verifiable Table-to-Text Generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, Franck Dernoncourt, Daniel Preoţiuc-Pietro, and Anastasia Shimorina (Eds.).

[7] Chengliang Chai, Nan Tang, Ju Fan, and Yuyu Luo. [n. d.]. Demystifying Artificial Intelligence for Data Preparation. In *Companion of the 2023 International Conference on Management of Data.*

[8] Chengliang Chai, Jiayi Wang, Yuyu Luo, Zeping Niu, and Guoliang Li. 2023. Data Management for Machine Learning: A Survey. *IEEE Transactions on Knowledge and Data Engineering* (2023).

[9] Sibei Chen, Ju Fan, Bin Wu, Nan Tang, Chao Deng, Pengyi Wang, Ye Li, Jian Tan, Feifei Li, Jingren Zhou, et al. 2025. Automatic database configuration debugging using retrieval-augmented language models. *Proceedings of the ACM on Management of Data* (2025).

[10] Zui Chen, Lei Cao, Sam Madden, Tim Kraska, Zeyuan Shang, Ju Fan, Nan Tang, Zihui Gu, Chunwei Liu, and Michael Cafarella. 2023. SEED: Domain-specific data curation with large language models. *arXiv preprint arXiv:2310.00749* (2023).

[11] Zhoujun Cheng, Tianbao Xie, Peng Shi, Chengzu Li, Rahul Nadkarni, Yushi Hu, Caiming Xiong, Dragomir Radev, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. 2023. Binding Language Models in Symbolic Languages. arXiv:2210.02875 [cs]

[12] Databricks. [n. d.]. *Assistant Data Science Agent.*

[13] Minghang Deng, Ashwin Ramachandran, Canwen Xu, Lanxiang Hu, Zhewei Yao, Anupam Datta, and Hao Zhang. [n. d.]. ReFoRCE: A Text-to-SQL Agent with Self-Refinement, Format Restriction, and Column Exploration. *arXiv preprint arXiv:2502.00675* ([n. d.]).

[14] Yuchen Duan, Zhe Chen, Yusong Hu, Weiyun Wang, Shenglong Ye, Botian Shi, Lewei Lu, Qibin Hou, Tong Lu, Hongsheng Li, Jifeng Dai, and Wenhai Wang. 2025. Docopilot: Improving Multimodal Models for Document-Level Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

[15] Alvaro AA Fernandes, Martin Koehler, and et al. 2023. Data preparation: A technological perspective and review. *SN Computer Science* (2023).

[16] Benjamin Feuer, Yurong Liu, Chinmay Hegde, and Juliana Freire. 2024. ArcheType: A Novel Framework for Open-Source Column Type Annotation Using Large Language Models. *Proceedings of the VLDB Endowment* (2024).

[17] Yanjie Fu, Dongjie Wang, Wangyang Ying, Xiangliang Zhang, Huan Liu, and Jian Pei. 2025. Autonomous Data Agents: A New Opportunity for Smart Data. *arXiv preprint arXiv:2509.18710* (2025).

[18] Victor Giannakouris and Immanuel Trummer. 2024. Dbg-pt: A large language model assisted query performance regression debugger. *Proceedings of the VLDB Endowment* (2024).

[19] Victor Giannakouris and Immanuel Trummer. 2025. λ-tune: Harnessing large language models for automated database system tuning. *Proceedings of the ACM on Management of Data* (2025).

[20] Google Cloud. [n. d.]. *BigQuery.*

[21] Jiawei Guan, Feng Zhang, Siqi Ma, Kuangyu Chen, Yihua Hu, Yuxing Chen, Anqun Pan, and Xiaoyong Du. 2023. Homomorphic Compression: Making Text Processing on Compression Unlimited. *Proc. ACM Manag. Data* 1, 4, Article 271 (Dec. 2023), 28 pages. doi:10.1145/3626765

[22] Sirui Hong, Yizhang Lin, and et al. Bang Liu. 2025. Data Interpreter: An LLM Agent for Data Science. In *Findings of the Association for Computational Linguistics.*

[23] Xinmei Huang, Haoyang Li, Jing Zhang, Xinxin Zhao, Zhiming Yao, Yiyan Li, Tieying Zhang, Jianjun Chen, Hong Chen, and Cuiping Li. 2025. E2Etune: End-to-end knob tuning via fine-tuned generative language model. *Proceedings of the VLDB Endowment* (2025).

[24] Jie Jiang, Haining Xie, Siqi Shen, Yu Shen, and et al. 2025. SiriusBI: A Comprehensive LLM-powered Solution for Data Analytics in Business Intelligence. *Proc. VLDB Endow.* (2025).

[25] Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. [n. d.]. Active Retrieval Augmented Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing.*

[26] Moe Kayali, Anton Lykov, Ilias Fountalis, Nikolaos Vasiloglou, Dan Olteanu, and Dan Suciu. 2023. CHORUS: foundation models for unified data discovery and exploration. *arXiv preprint arXiv:2306.09610* (2023).

[27] Keti Korini and Christian Bizer. 2025. Evaluating knowledge generation and self-refinement strategies for llm-based column type annotation. *arXiv preprint arXiv:2503.02718* (2025).

[28] Jiale Lao, Yibo Wang, Yufei Li, Jianping Wang, Yunjia Zhang, Zhiyuan Cheng, Wanghu Chen, Mingjie Tang, and Jianguo Wang. 2024. GPTuner: A Manual-Reading Database Tuning System via GPT-Guided Bayesian Optimization. *Proc. VLDB Endow.* (2024).

[29] Boyan Li, Chong Chen, Zhujun Xue, Yinan Mei, and Yuyu Luo. 2025. DeepEye-SQL: A Software-Engineering-Inspired Text-to-SQL Framework. *arXiv preprint arXiv:2510.17586* (2025).

[30] Boyan Li, Yuyu Luo, Chengliang Chai, Guoliang Li, and Nan Tang. 2024. The Dawn of Natural Language to SQL: Are We Fully Ready? *Proc. VLDB Endow.* (2024).

[31] Boyan Li, Yiran Peng, Yupeng Xie, Sirong Lu, Yizhang Zhu, Xing Mu, Xinyu Liu, and Yuyu Luo. 2026. DeepEye: A Steerable Self-driving Data Agent System. In *Companion of the 2026 International Conference on Management of Data (SIGMOD Companion '26).* ACM, Bengaluru, India. doi:10.1145/3788853.3801612

[32] Boyan Li, Jiayi Zhang, Ju Fan, Yanwei Xu, Chong Chen, Nan Tang, and Yuyu Luo. 2025. Alpha-SQL: Zero-Shot Text-to-SQL using Monte Carlo Tree Search. In *Forty-second International Conference on Machine Learning.*

[33] Changlun Li, Yao SHI, Chen Wang, Qiqi Duan, Runke RUAN, Weijie Huang, Haonan Long, Lijun Huang, Nan Tang, and Yuyu Luo. 2025. Time Travel is Cheating: Going Live with DeepFund for Real-Time Fund Investment Benchmarking. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track.* https://openreview.net/forum?id=SXADEhZ0sl

[34] Changlun Li, Chenyu Yang, Yuyu Luo, Ju Fan, and Nan Tang. 2025. Weak-to-Strong Prompts with Lightweight-to-Powerful LLMs for High-Accuracy, Low-Cost, and Explainable Data Transformation. *Proceedings of the VLDB Endowment* (2025).

[35] Guoliang Li, Jiayi Wang, Chenyang Zhang, and Jiannan Wang. 2025. Data+AI: LLM4Data and Data4LLM. In *Companion of the 2025 International Conference on Management of Data.*

[36] Guoliang Li, Xuanhe Zhou, and Xinyang Zhao. 2024. LLM for data management. *Proceedings of the VLDB Endowment* (2024).

[37] Peng Li, Yeye He, Dror Yashar, Weiwei Cui, Song Ge, Haidong Zhang, Danielle Rifinski Fainman, Dongmei Zhang, and Surajit Chaudhuri. 2024. Table-gpt: Table fine-tuned gpt for diverse table tasks. *Proceedings of the ACM on Management of Data* (2024).

[38] Shuaimin Li, Xuanang Chen, Yuanfeng Song, Yunze Song, Chen Jason Zhang, Fei Hao, and Lei Chen. 2025. prompt4vis: prompting large language models with example mining for tabular data visualization. *The VLDB Journal* (2025).

[39] Yuchen Li, Kai Wang, Zhiqiang Sun, et al. 2025. Automatic Database Description Generation for Text-to-SQL. *arXiv preprint arXiv:2502.20657* (2025).

[40] Zhaodonghui Li, Haitao Yuan, Huiming Wang, Gao Cong, and Lidong Bing. 2024. LLM-R2: A Large Language Model Enhanced Rule-Based Rewrite System for Boosting Query Efficiency. *Proceedings of the VLDB Endowment* (2024).

[41] Xiaotian Lin, Yanlin Qi, Yizhang Zhu, Themis Palpanas, Chengliang Chai, Nan Tang, and Yuyu Luo. 2026. LEAD: Iterative Data Selection for Efficient LLM Instruction Tuning. *Proc. VLDB Endow.* 19, 3 (March 2026), 426–439. doi:10.14778/3778092.3778103

[42] Bang Liu, Xinfeng Li, Jiayi Zhang, Jinlin Wang, Tanjin He, Sirui Hong, Hongzhang Liu, Shaokun Zhang, Kaitao Song, Kunlun Zhu, et al. 2025. Advances and challenges in foundation agents: From brain-inspired intelligence to evolutionary, collaborative, and safe systems. *arXiv preprint arXiv:2504.01990* (2025).

[43] Xinyu Liu, Shuyu Shen, Boyan Li, Peixian Ma, Runzhi Jiang, Yuxin Zhang, Ju Fan, Guoliang Li, Nan Tang, and Yuyu Luo. 2025. A Survey of Text-to-SQL in the Era of LLMs: Where are we, and where are we going? *IEEE Transactions on Knowledge and Data Engineering* (2025).

[44] Xinyu Liu, Shuyu Shen, Boyan Li, Nan Tang, and Yuyu Luo. 2025. Nl2sql-bugs: A benchmark for detecting semantic errors in nl2sql translation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2.*

[45] Tianqi Luo, Chuhan Huang, Leixian Shen, Boyan Li, Shuyu Shen, Wei Zeng, Nan Tang, and Yuyu Luo. 2025. nvBench 2.0: Resolving Ambiguity in Text-to-Visualization through Stepwise Reasoning. arXiv:2503.12880 [cs.CL]

[46] Yuyu Luo, Guoliang Li, Ju Fan, Chengliang Chai, and Nan Tang. 2025. Natural language to sql: State of the art and open problems. *Proceedings of the VLDB Endowment* (2025).

[47] Yuyu Luo, Xuedi Qin, Nan Tang, and Guoliang Li. 2018. DeepEye: Towards Automatic Data Visualization. In *ICDE.* IEEE Computer Society.

[48] Yuyu Luo, Nan Tang, Guoliang Li, Chengliang Chai, Wenbo Li, and Xuedi Qin. 2021. Synthesizing Natural Language to Visualization (NL2VIS) Benchmarks from NL2SQL Benchmarks. In *Proceedings of the 2021 International Conference on Management of Data (SIGMOD '21).*

[49] Yuyu Luo, Nan Tang, Guoliang Li, Jiawei Tang, Chengliang Chai, and Xuedi Qin. 2022. Natural Language to Visualization by Neural Machine Translation. *IEEE Transactions on Visualization and Computer Graphics* (2022).

[50] Yuyu Luo, Yihui Zhou, Nan Tang, Guoliang Li, Chengliang Chai, and Leixian Shen. 2023. Learned Data-aware Image Representations of Line Charts for Similarity Search. *Proc. ACM Manag. Data* (2023).

[51] Paula Maddigan and Teo Susnjak. 2023. Chat2VIS: Generating Data Visualizations via Natural Language Using ChatGPT, Codex and GPT-3 Large Language Models. *IEEE Access* (oct 2023).

[52] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196* (2024).

[53] Zan Ahmad Naeem, Mohammad Shahmeer Ahmad, Mohamed Eltabakh, Mourad Ouzzani, and Nan Tang. 2024. RetClean: Retrieval-Based Data Cleaning Using LLMs and Data Lakes. *Proc. VLDB Endow.* (2024).

[54] Avanika Narayan, Ines Chami, Laurel Orr, and Christopher Ré. 2022. Can Foundation Models Wrangle Your Data? *Proceedings of the VLDB Endowment* (2022).

[55] Geliang Ouyang, Jingyao Chen, Zhihe Nie, Yi Gui, Yao Wan, Hongyu Zhang, and Dongping Chen. 2025. nvAgent: Automated Data Visualization from Natural Language via Collaborative Agent Workflow. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics.*

[56] Fatemeh Pesaran Zadeh, Juyeon Kim, Jin-Hwa Kim, and Gunhee Kim. [n. d.]. Text2Chart31: Instruction Tuning for Chart Generation with Automatic Feedback. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing.*

[57] Mohammadreza Pourreza, Hailong Li, Ruoxi Sun, Yeounoh Chung, Shayan Talaei, Gaurav Tarlok Kakkar, Yu Gan, Amin Saberi, Fatma Ozcan, and Sercan O Arik. 2025. CHASE-SQL: Multi-Path Reasoning and Preference Optimized Candidate Selection in Text-to-SQL. In *The Thirteenth International Conference on Learning Representations.*

[58] Mohammadreza Pourreza and Davood Rafiei. 2023. DIN-SQL: Decomposed In-Context Learning of Text-to-SQL with Self-Correction. In *Advances in Neural Information Processing Systems 36*, Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (Eds.).

[59] Danrui Qi, Zhengjie Miao, and Jiannan Wang. 2025. CleanAgent: Automating Data Standardization with LLM-based Agents. arXiv:2403.08291

[60] Zhangcheng Qiang, Weiqing Wang, and Kerry Taylor. 2024. Agent-OM: Leveraging LLM Agents for Ontology Matching. *Proceedings of the VLDB Endowment* (2024).

[61] Jon Saad-Falcon, Joe Barrow, Alexa Siu, Ani Nenkova, Seunghyun Yoon, Ryan A. Rossi, and Franck Dernoncourt. [n. d.]. PDFTriage: Question Answering over Long, Structured Documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track.*

[62] Elisabeth Shi, Tom Michael Gasser, Andre Seeck, and Rico Auerswald. 2020. The Principles of Operation Framework: A Comprehensive Classification Concept for Automated Driving Functions. *SAE International Journal of Connected and Automated Vehicles* (2020).

[63] Zhihao Shuai, Boyan Li, Siyu Yan, Yuyu Luo, and Weikai Yang. 2025. Deepvis: Bridging natural language and data visualization through step-wise reasoning. *arXiv preprint arXiv:2508.01700* (2025).

[64] Snowflake. [n. d.]. *Cortex Agents.*

[65] Yuyang Song, Hanxu Yan, Jiale Lao, Yibo Wang, Yufei Li, Yuanchun Zhou, Jianguo Wang, and Mingjie Tang. 2025. QUITE: A Query Rewrite System Beyond Rules with LLM Agents. *arXiv preprint arXiv:2506.07675* (2025).

[66] Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. 2024. Table Meets LLM: Can Large Language Models Understand Structured Table Data? A Benchmark and Empirical Study. arXiv:2305.13062 [cs]

[67] Nicole Sultanum and Arjun Srinivasan. 2023. DATATALES: Investigating the use of Large Language Models for Authoring Data-Driven Articles. In *2023 IEEE Visualization and Visual Analytics (VIS).*

[68] Ji Sun, Guoliang Li, Peiyao Zhou, Yihui Ma, Jingzhe Xu, and Yuan Li. 2025. AgenticData: An Agentic Data Analytics System for Heterogeneous Data. *CoRR* (2025).

[69] Wenwen Sun, Zhicheng Pan, Zirui Hu, Yu Liu, Chengcheng Yang, Rong Zhang, and Xuan Zhou. [n. d.]. Rabbit: Retrieval-Augmented Generation Enables Better Automatic Database Knob Tuning. In *2025 IEEE 41st International Conference on Data Engineering (ICDE).* IEEE.

[70] Zhaoyan Sun, Jiayi Wang, Xinyang Zhao, Jiachi Wang, and Guoliang Li. 2025. Data Agent: A Holistic Architecture for Orchestrating Data+ AI Ecosystems. *arXiv preprint arXiv:2507.01599* (2025).

[71] Zhaoyan Sun, Xuanhe Zhou, Guoliang Li, Xiang Yu, Jianhua Feng, and Zhangm Yong. 2025. R-Bot: An LLM-based Query Rewrite System. *Proceedings of the VLDB Endowment* (2025).

[72] Manan Suri, Puneet Mathur, Franck Dernoncourt, Kanika Goswami, Ryan A. Rossi, and Dinesh Manocha. [n. d.]. VisDoM: Multi-Document QA with Visually Rich Elements Using Multimodal Retrieval-Augmented Generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies.*

[73] Manan Suri, Puneet Mathur, Nedim Lipka, Franck Dernoncourt, Ryan A. Rossi, and Dinesh Manocha. 2025. ChartLens: Fine-grained Visual Attribution in Charts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics.*

[74] TabTab AI. [n. d.]. *TabTab.*

[75] Zirui Tang, Weizheng Wang, Zihang Zhou, Yang Jiao, Bangrui Xu, Boyu Niu, Xuanhe Zhou, Guoliang Li, Yeye He, Wei Zhou, et al. 2025. LLM/Agent-as-Data-Analyst: A Survey. *arXiv preprint arXiv:2509.23988* (2025).

[76] Fengwei Teng, Zhaoyang Yu, Quan Shi, Jiayi Zhang, Chenglin Wu, and Yuyu Luo. 2025. Atom of Thoughts for Markov LLM Test-Time Scaling. *CoRR* (2025).

[77] Fen Wang, Bomiao Wang, Xueli Shu, Zhen Liu, Zekai Shao, Chao Liu, and Siming Chen. 2025. ChartInsighter: An Approach for Mitigating Hallucination in Time-series Chart Summary Generation with A Benchmark Dataset. arXiv:2501.09349 [cs.CL]

[78] Jiayi Wang and Guoliang Li. 2025. AOP: Automated and interactive llm pipeline orchestration for answering complex queries. CIDR.

[79] Jiayi Wang, Guoliang Li, and Jianhua Feng. 2025. iDataLake: An llm-powered analytics system on data lakes. *Data Engineering* (2025).

[80] Yuhao Wang, Ruiyang Ren, Junyi Li, Xin Zhao, Jing Liu, and Ji-Rong Wen. [n. d.]. REAR: A Relevance-Aware Retrieval-Augmented Framework for Open-Domain Question Answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing.*

[81] Yifan Wu, Yiran Peng, Yiyu Chen, Jianhao Ruan, Zijie Zhuang, Cheng Yang, Jiayi Zhang, Man Chen, Yenchi Tseng, Zhaoyang Yu, Liang Chen, Yuyao Zhai, Bang Liu, Chenglin Wu, and Yuyu Luo. 2026. AutoWebWorld: Synthesizing Infinite Verifiable Web Environments via Finite State Machines. arXiv:2602.14296 [cs.AI] https://arxiv.org/abs/2602.14296

[82] Yifan Wu, Lutao Yan, Leixian Shen, Yunhai Wang, Nan Tang, and Yuyu Luo. 2024. ChartInsights: Evaluating Multimodal Large Language Models for Low-Level Chart Question Answering. In *EMNLP (Findings).* Association for Computational Linguistics, 12174–12200.

[83] Yupeng Xie, Yuyu Luo, Guoliang Li, and Nan Tang. 2024. HAIChart: Human and AI Paired Visualization System. *Proceedings of the VLDB Endowment* (2024).

[84] Yupeng Xie, Zhiyang Zhang, Yifan Wu, Sirong Lu, Jiayi Zhang, Zhaoyang Yu, Jinlin Wang, Sirui Hong, Bang Liu, Chenglin Wu, and Yuyu Luo. 2025. VisJudge-Bench: Aesthetics and Quality Assessment of Visualizations. *CoRR* abs/2510.22373 (2025).

[85] Wenyi Xu, Yuren Mao, Xiaolu Zhang, Chao Zhang, Xuemei Dong, Mengfei Zhang, and Yunjun Gao. 2025. DAgent: A Relational Database-Driven Data Analysis Report Generation Agent. arXiv:2503.13269 [cs.DB]

[86] Zihan Yan, Rui Xi, and Mengshu Hou. 2025. MCTuner: Spatial Decomposition-Enhanced Database Tuning via LLM-Guided Exploration. *arXiv preprint arXiv:2509.06298* (2025).

[87] Chenyu Yang, Yuyu Luo, Chuanxuan Cui, Ju Fan, Chengliang Chai, and Nan Tang. 2025. Data Imputation with Limited Data Redundancy Using Data Lakes. *Proc. VLDB Endow.* (2025). doi:10.14778/3748191.3748200

[88] Zhaorui Yang, Bo Pan, Han Wang, Yiyao Wang, Xingyu Liu, Minfeng Zhu, Bo Zhang, and Wei Chen. 2025. Multimodal DeepResearcher: Generating Text-Chart Interleaved Reports From Scratch with Agentic Framework. arXiv:2506.02454 [cs.CL]

[89] Zhiyu Yang, Zihan Zhou, Shuo Wang, Xin Cong, Xu Han, Yukun Yan, Zhenghao Liu, Zhixing Tan, Pengyuan Liu, Dong Yu, Zhiyuan Liu, Xiaodong Shi, and Maosong Sun. [n. d.]. MatPlotAgent: Method and Evaluation for LLM-Based Agentic Scientific Data Visualization. In *Findings of the Association for Computational Linguistics 2024.*

[90] Yunhu Ye, Binyuan Hui, Min Yang, Binhua Li, Fei Huang, and Yongbin Li. 2023. Large Language Models Are Versatile Decomposers: Decomposing Evidence and Questions for Table-based Reasoning. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval.*

[91] Peiying Yu, Guoxin Chen, and Jingjing Wang. 2025. Table-Critic: A Multi-Agent Framework for Collaborative Criticism and Refinement in Table Reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics.*

[92] Feng Zhang, Jidong Zhai, Xipeng Shen, Dalin Wang, Zheng Chen, Onur Mutlu, Wenguang Chen, and Xiaoyong Du. 2021. TADOC: Text analytics directly on compression. *The VLDB Journal* 30, 2 (2021), 163–188.

[93] Huan Zhang, Yizhan Li, Wenhao Huang, Ziyu Hou, Yu Song, Xuye Liu, Farshid Effaty, Jinya Jiang, Sifan Wu, Qianggang Ding, Izumi Takahara, Leonard R. MacGillivray, Teruyasu Mizoguchi, Tianshu Yu, Lizi Liao, Yuyu Luo, Yu Rong, Jia Li, Ying Diao, Heng Ji, and Bang Liu. 2026. Towards Agentic Intelligence for Materials Science. arXiv:2602.00169 [cond-mat.mtrl-sci] https://arxiv.org/abs/2602.00169

[94] Haoxiang Zhang, Yurong Liu, Aécio Santos, Juliana Freire, et al. 2025. Autoddg: Automated dataset description generation using large language models. *arXiv preprint arXiv:2502.01050* (2025).

[95] Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xionghui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, et al. 2024. Aflow: Automating agentic workflow generation. *arXiv preprint arXiv:2410.10762* (2024).

[96] Shaolei Zhang, Ju Fan, Meihao Fan, Guoliang Li, and Xiaoyong Du. 2025. DeepAnalyze: Agentic Large Language Models for Autonomous Data Science. arXiv:2510.16872 [cs.AI]

[97] Xiang Zhang, Hongming Xu, Le Zhou, Wei Zhou, Xuanhe Zhou, Guoliang Li, Yuyu Luo, Changdong Liu, Guorun Chen, Jiang Liao, and Fan Wu. 2026. Dial: A Knowledge-Grounded Dialect-Specific NL2SQL System. arXiv:2603.07449 [cs.DB] https://arxiv.org/abs/2603.07449

[98] Yuxin Zhang, Meihao Fan, Ju Fan, Mingyang Yi, Yuyu Luo, Jian Tan, and Guoliang Li. 2025. Reward-SQL: Boosting Text-to-SQL via Stepwise Reasoning and Process-Supervised Rewards. arXiv:2505.04671 [cs.CL] https://arxiv.org/abs/2505.04671

[99] Zhengxuan Zhang, Zhuowen Liang, Yin Wu, Teng Lin, Yuyu Luo, and Nan Tang. 2025. DataPuzzle: Breaking Free from the Hallucinated Promise of LLMs in Data Analysis. *arXiv preprint arXiv:2504.10036* (2025).

[100] Xinyang Zhao, Xuanhe Zhou, and Guoliang Li. 2023. Automatic database knob tuning: A survey. *IEEE Transactions on Knowledge and Data Engineering* (2023).

[101] Wei Zhou, Ji Sun, Xuanhe Zhou, Guoliang Li, Luyang Liu, Hao Wu, and Tianyuan Wang. 2025. GaussMaster: An LLM-based Database Copilot System. *arXiv preprint arXiv:2506.23322* (2025).

[102] Wei Zhou, Jun Zhou, Haoyu Wang, Zhenghao Li, Qikang He, Shaokun Han, Guoliang Li, Xuanhe Zhou, Yeye He, Chunwei Liu, Zirui Tang, Bin Wang, Shen Tang, Kai Zuo, Yuyu Luo, Zhenzhe Zheng, Conghui He, Jingren Zhou, and Fan Wu. 2026. Can LLMs Clean Up Your Mess? A Survey of Application-Ready Data Preparation with LLMs. arXiv:2601.17058 [cs.DB] https://arxiv.org/abs/2601.17058

[103] Xuanhe Zhou, Junxuan He, Wei Zhou, Haodong Chen, Zirui Tang, Haoyu Zhao, Xin Tong, Guoliang Li, Youmin Chen, Jun Zhou, et al. 2025. A Survey of LLM× DATA. *arXiv preprint arXiv:2505.18458* (2025).

[104] Xuanhe Zhou, Guoliang Li, and Zhiyuan Liu. 2023. LLM as DBA. *arXiv preprint arXiv:2308.05481* (2023).

[105] Xuanhe Zhou, Guoliang Li, Zhaoyan Sun, Zhiyuan Liu, Weize Chen, Jianming Wu, Jiesi Liu, Ruohang Feng, and Guoyang Zeng. 2024. D-Bot: Database Diagnosis System using Large Language Models. *Proceedings of the VLDB Endowment* (2024).

[106] Yizhang Zhu, Shiyin Du, Boyan Li, Yuyu Luo, and Nan Tang. 2024. Are large language models good statisticians? *Advances in Neural Information Processing Systems* (2024).

[107] Yizhang Zhu, Runzhi JIANG, Boyan Li, Nan Tang, and Yuyu Luo. 2025. EllieSQL: Cost-Efficient Text-to-SQL with Complexity-Aware Routing. In *Second Conference on Language Modeling*.

[108] 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, Fugee Tsung, Jiannan Wang, Chenglin Wu, Yanwei Xu, Shaolei Zhang, Yong Zhang, Xuanhe Zhou, Guoliang Li, and Yuyu Luo. 2025. A Survey of Data Agents: Emerging Paradigm or Overstated Hype? arXiv:2510.23587 [cs.DB] https://arxiv.org/abs/2510.23587