THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

DA LAB

# A Survey of Data Agents: Emerging Paradigm or Overstated Hype?

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*, Yuyu LUO*

https://luoyuyu.vip
https://github.com/HKUSTDial/awesome-data-agents

# Outline

- Introduction
- Hierarchical Taxonomy for Data Agents
- L0: Manual Labor in Early Ages
- L1: Preliminary Assistance
- L2: Perceive the Environment
- L3: Striving for Autonomous Data Agents
- L4-L5: Vision of Proactive and Generative Data Agents
- Conclusion

# The Dawn of Data Agents

- Tackling data-related tasks can be demanding
- Long-standing aspiration in data science and analytics: *developing an intelligent agent capable of autonomously managing, preparing, and analyzing data to deliver trustworthy insights with minimal human intervention.*

Data Management

Data Preparation

Data Analysis

Time-consuming

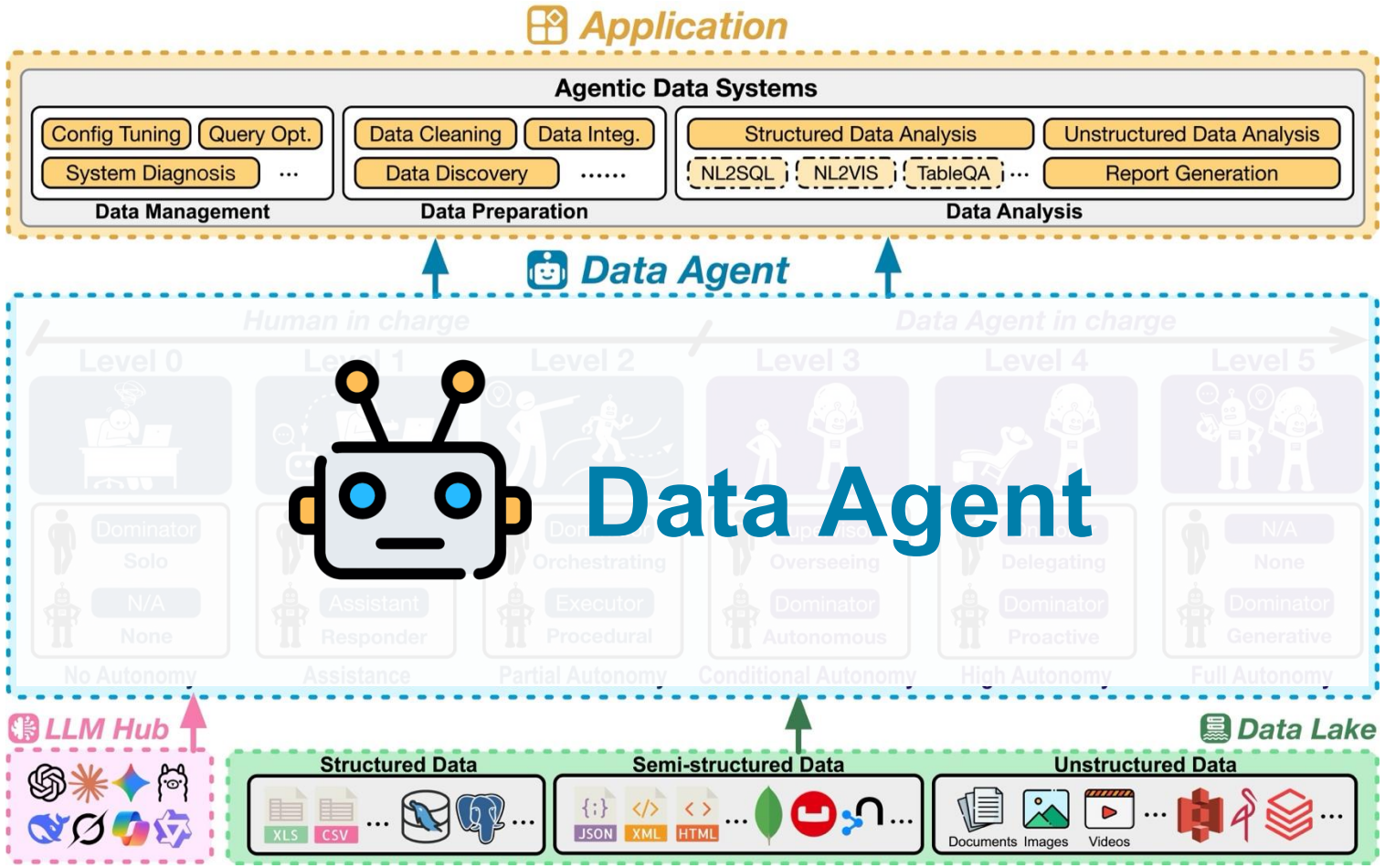
Labor-intensive

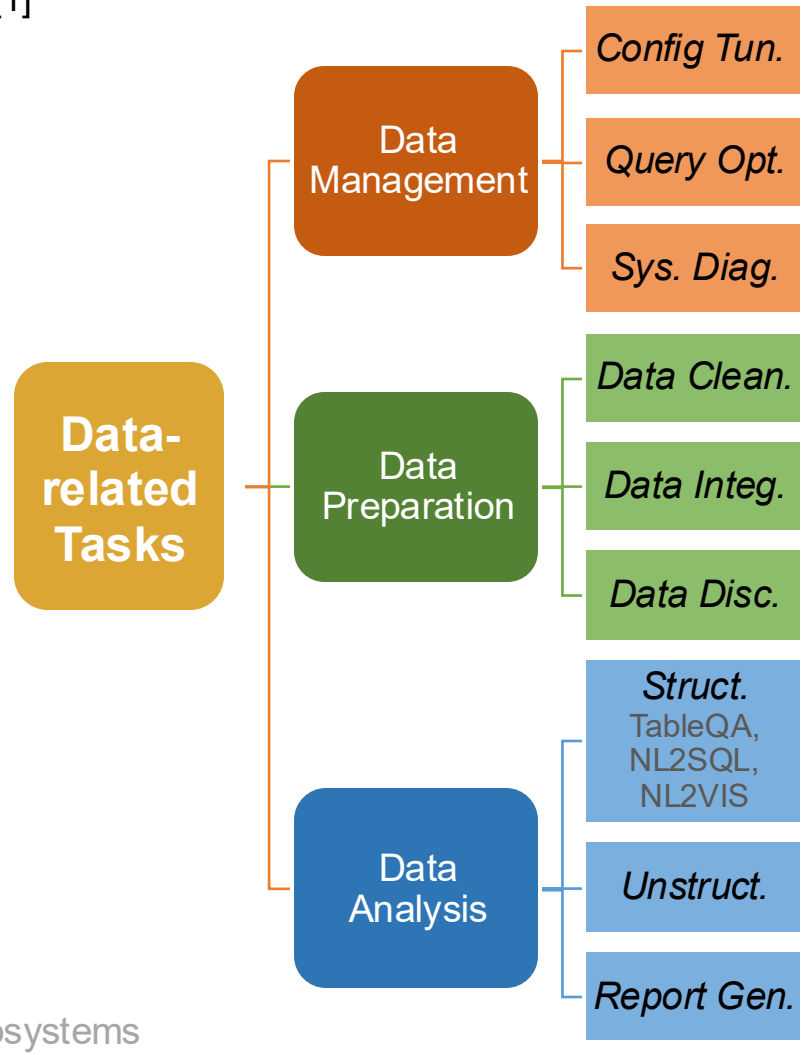
Knowledge-heavy

LLMs and LLM agents are bringing us closer to this vision

# The Dawn of Data Agents

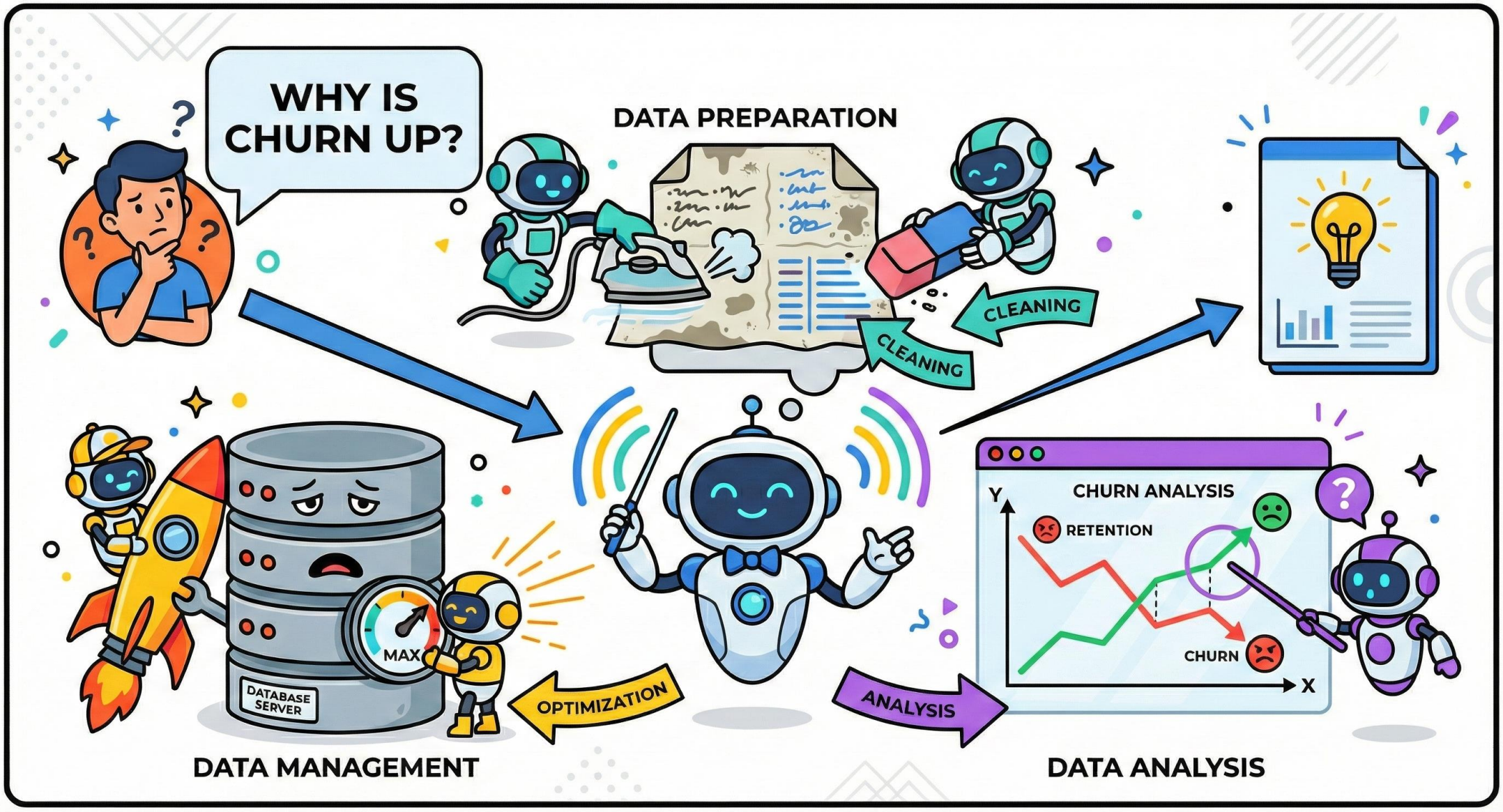
- Data Agent: a comprehensive architecture designed to orchestrate the **Data** + **AI** ecosystem, which autonomously addresses a wide spectrum of **data-related task** [1]




[1] Sun Z, Wang J, Zhao X, et al. Data Agent: A Holistic Architecture for Orchestrating Data+ AI Ecosystems

# The Dawn of Data Agents: An Example



# Data Agents vs. General LLM Agents

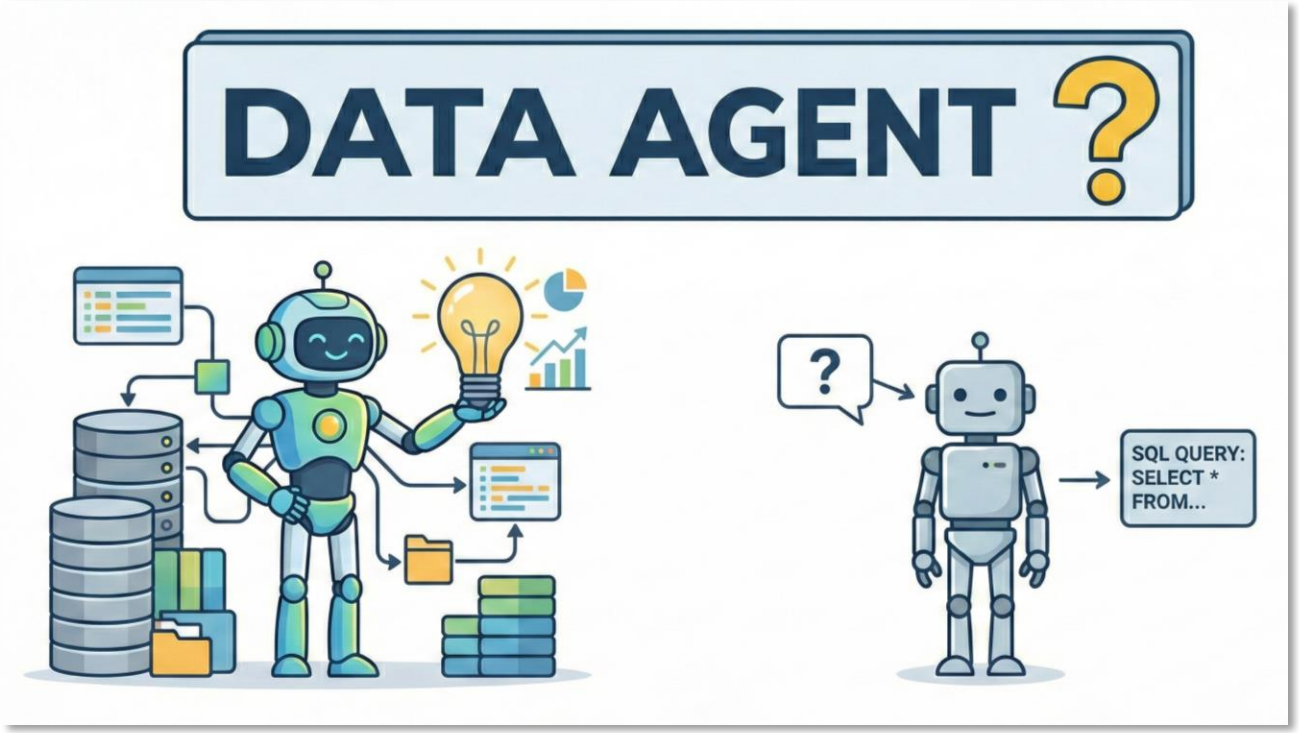
Formally, a data agent $A$ operates on raw data $D$ within an environment $E$ (e.g., DBMS, code interpreters, APIs, etc.), utilizing LLMs $M$, ultimately producing an output $O$ to tackle the data-related task $T$ :

$$A: (T, D, E, M) \rightarrow O$$

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

# The Terminological Ambiguity of Data Agents

- The term “Data Agent” is applied inconsistently:
  - Sophisticated agentic data systems to autonomously interact with data lakes, invoke external tools, orchestrate and optimize tailored pipelines for complex data-related tasks
  - More rudimentary, narrowly scoped systems acting as simple query responders




Conflate systems of different autonomy, reliability, and complexity under a imprecisely defined umbrella term.

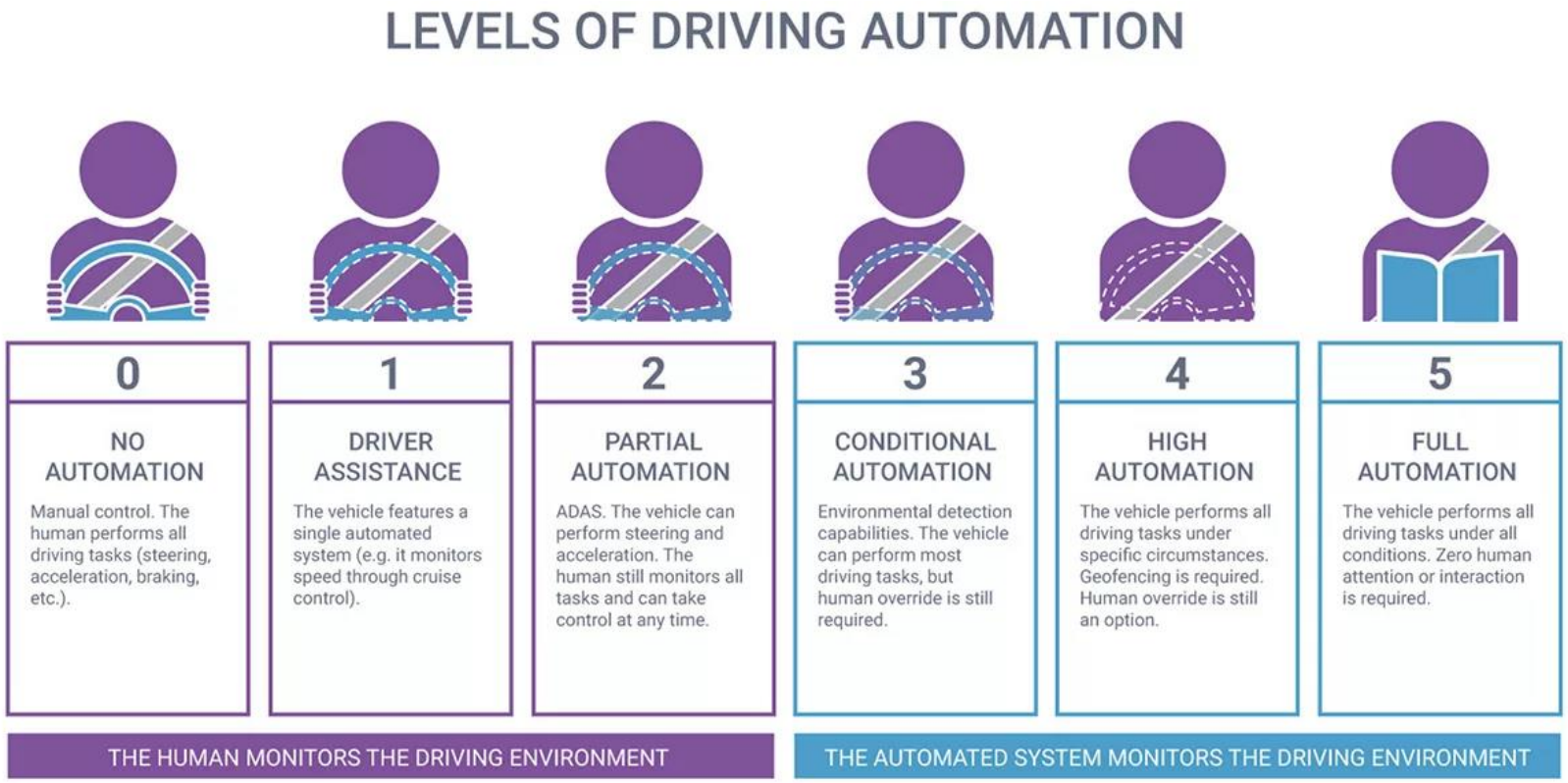
- **User-Side Risk:** User expectation mismatch
- **Governance Risk:** Unclear accountability
- **Industry-Side Risk:** Exaggeration and hype

# Hierarchical Taxonomy for Data Agents

## Precedent in Self-Driving

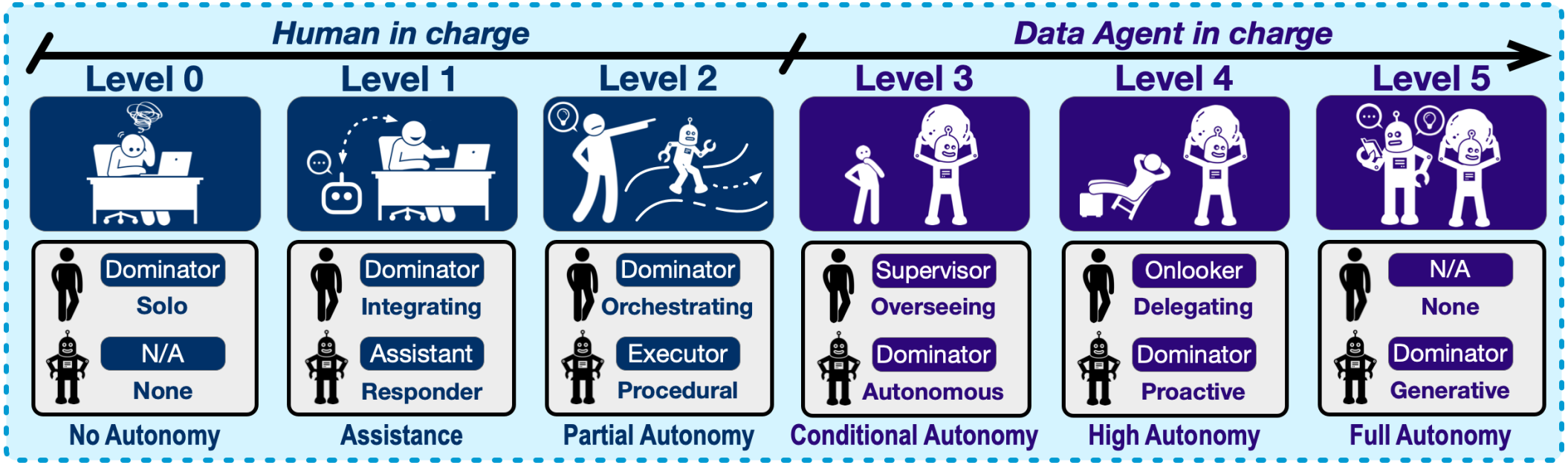
- Such a terminological ambiguity is not unprecedented:

  Automotive industry and driving automation community had encountered similar challenges
- SAE introduced the J3016 standard, a six-level taxonomy for driving automation




https://www.synopsys.com/blogs/chip-design/autonomous-driving-levels.html

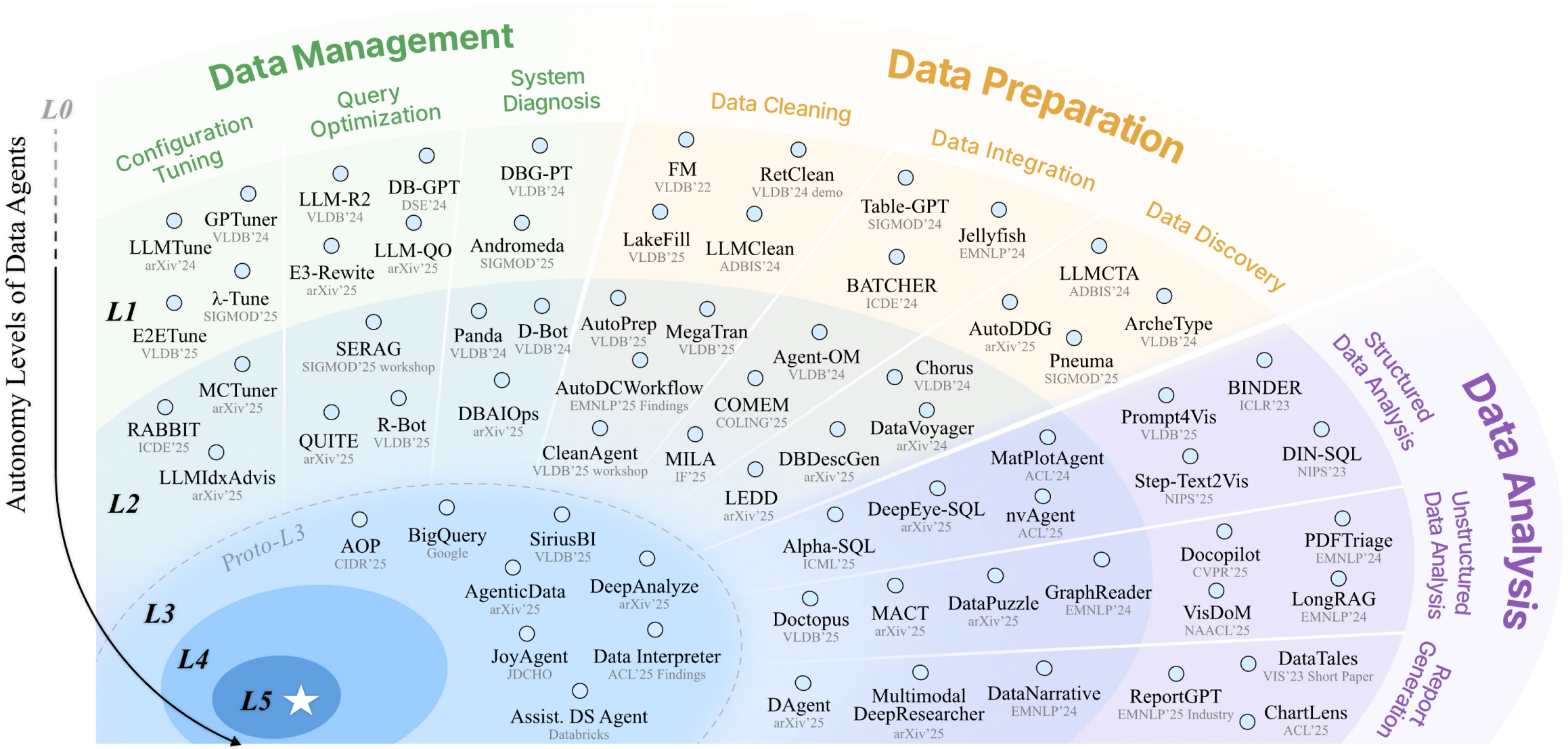
## We Advocate a Hierarchical Taxonomy for Data Agents




- Map the progressive transitions of dominance and responsibility in data-related tasks from human to data agent as autonomy increases from L0 to L5
- Unified framework to compare existing works, delineating capability boundaries, and clarifying accountability, enabling practitioners to align expectations and intervention with autonomy levels.
- We will elaborate on the formal definition for each level in the following

# Hierarchical Taxonomy for Data Agents

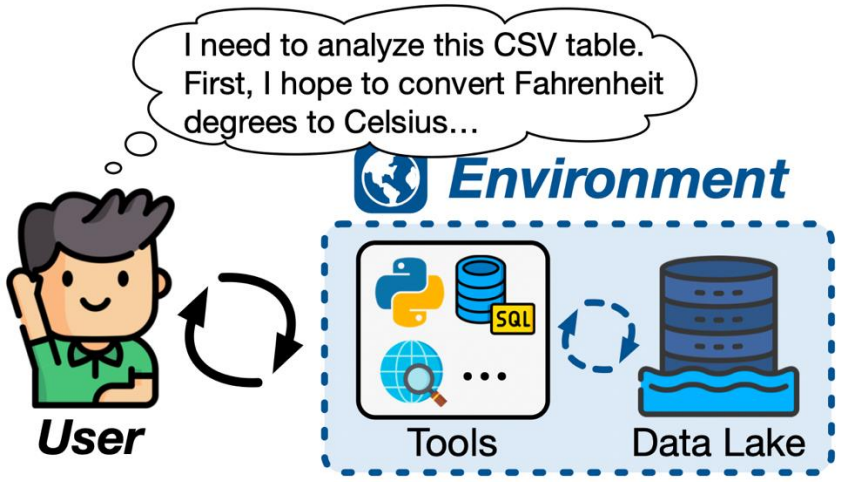
## Structured Review Through this Lens




Check our full paper list: https://github.com/HKUSTDial/awesome-data-agents

# L0: Manual Labor in Early Ages

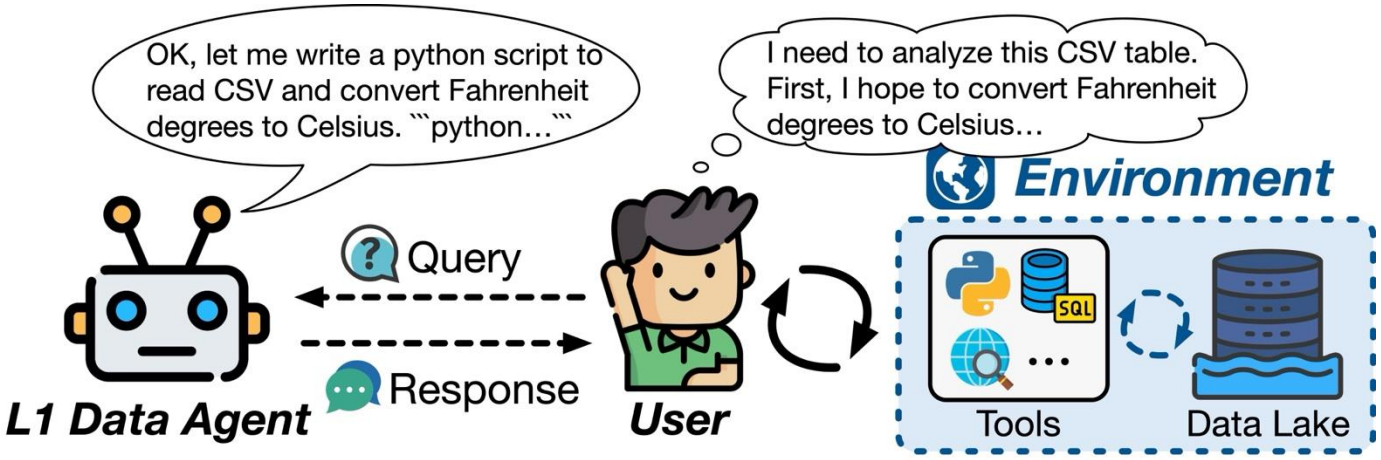
## Human-driven Data Management, Preparation, Analysis

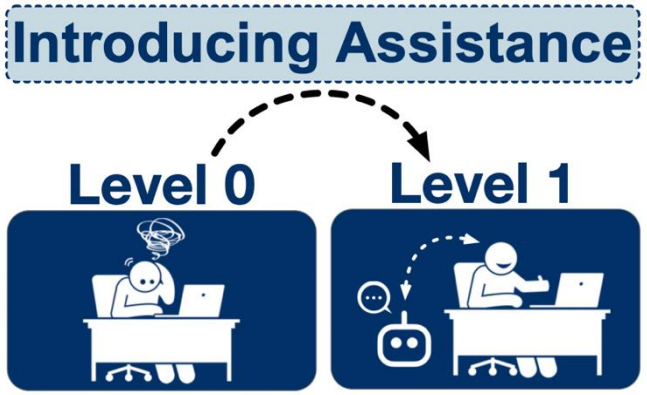



- Conventionally, all data management, preparation, and analysis tasks are performed entirely by humans without intelligent assistance.
- Formally, the human $H$ is responsible for the entire process, orchestrating ($\pi_H$) pipeline $P$ and executing ($\epsilon_H$), while the data agent $A$ is uninvolved yet:

$$H: \pi_H(T, D, E) \rightarrow P; \epsilon_H(P, D, E) \rightarrow O$$
$$A: \emptyset$$

# Definition for L1 Data Agents (Assistance)



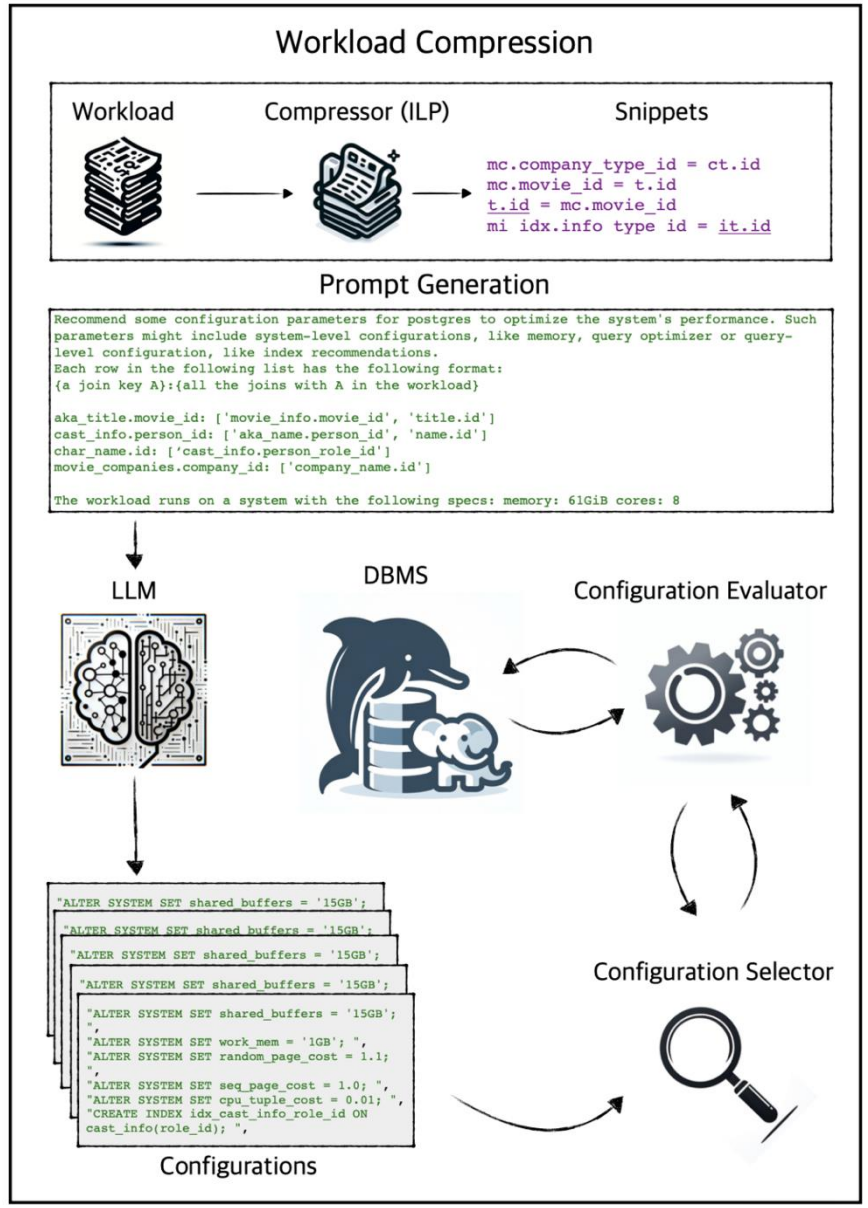




- Align with the early wave of LLM assistants
- Prompt-response paradigm: data agents act as nascent, stateless query-responsive assistants
- Incapable of perceiving and interacting with the environment
- Formally, the human $H$ remains responsible for both pipeline orchestration ($\pi_H$) and execution ($\epsilon_H$), while data agent $A$ can respond $r$ upon human query $q$ for assistance

$$H: \pi_H(T, D, E) \rightarrow P; \epsilon_H(P, D, E, r) \rightarrow O.$$
$$A: (q, M) \rightarrow r$$

## Config Tuning: $\lambda$-Tune (SIGMOD 2025)

- Construct prompt templates with information of OLAP workload, hardware specification, target database system
- Prompt LLMs to generate/recommend multiple database configuration candidates (SQL commands)
- Select the best configuration from the candidates

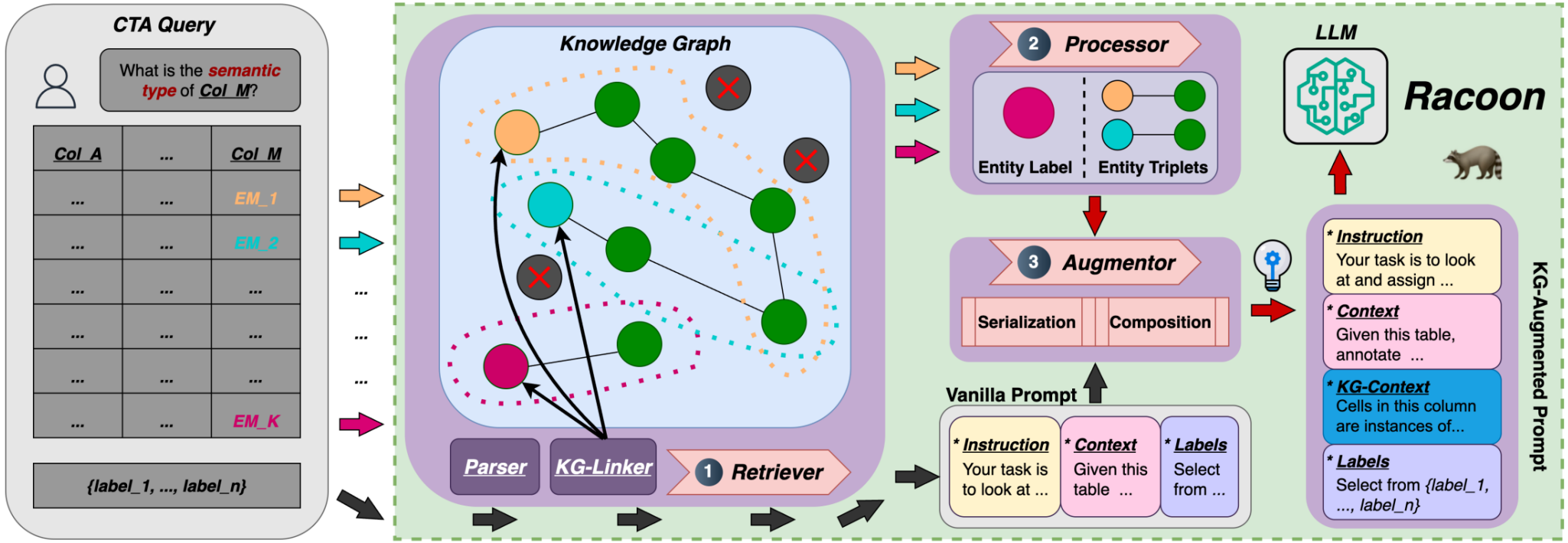





Giannakouris V, Trummer I. λ-tune: Harnessing large language models for automated database system tuning[J]. SIGMOD, 2025

# Data Discovery: RACOON (TRL@NeurIPS 2024)

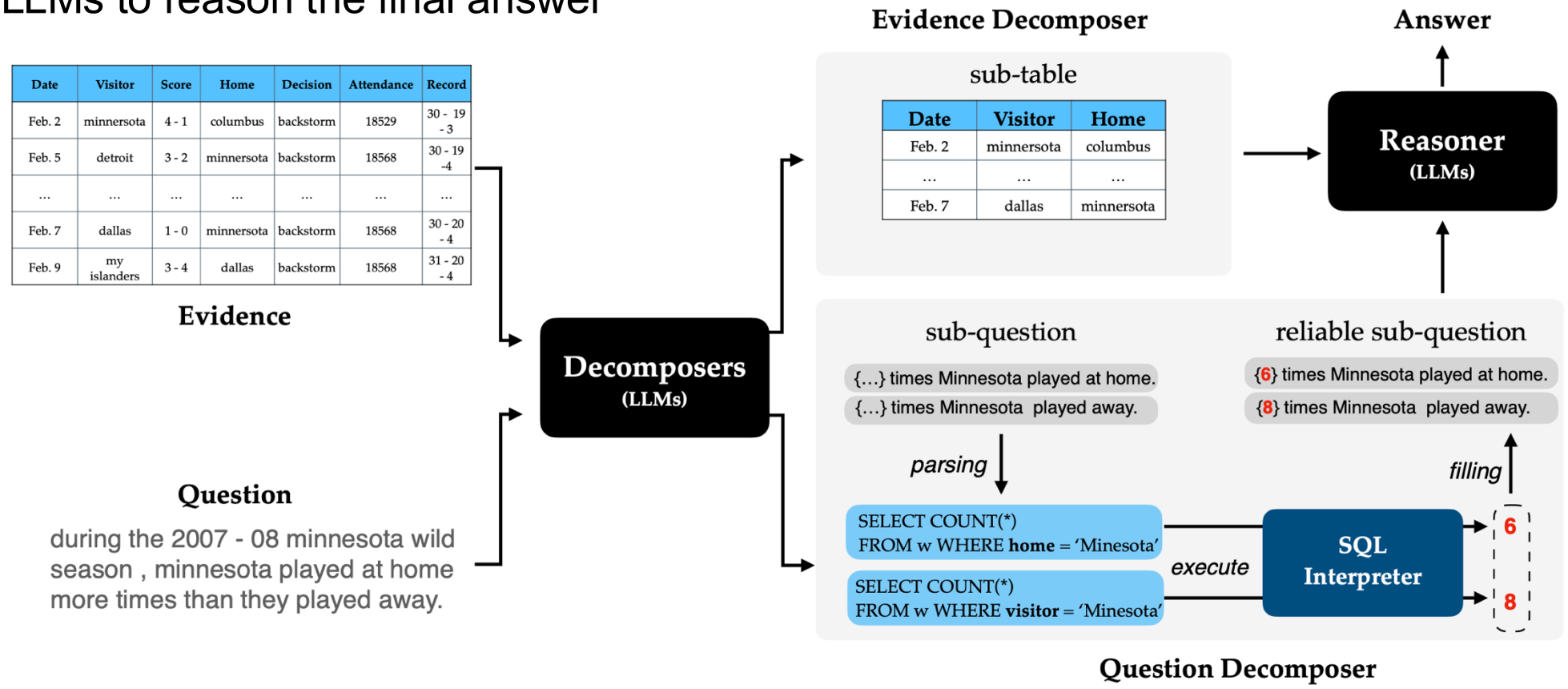
- Retrieval matching entities and relevant knowledge in the accompanying knowledge graph
- Construct contextual information based on retrieved knowledge
- Augment the prompt to enhance the performance of column-type annotation




Wei L L, Xiao G, Balazinska M. RACOON: An LLM-based Framework for Retrieval-Augmented Column Type Annotation with a Knowledge Graph[C]. NeurIPS 2024 Third Table Representation Learning Workshop.

# Structured Data Analysis: Dater (SIGIR 2023)

- Prompt LLMs to decompose huge tables into sub-tables, questions into sub-questions
- Tackle sub-questions using intermediate SQLs generated by LLMs
- Deploy LLMs to reason the final answer




Ye Y, Hui B, Yang M, et al. Large language models are versatile decomposers: Decomposing evidence and questions for table-based reasoning[C]. SIGIR, 2023

# L1: Preliminary Assistance (Progress and Limitations)

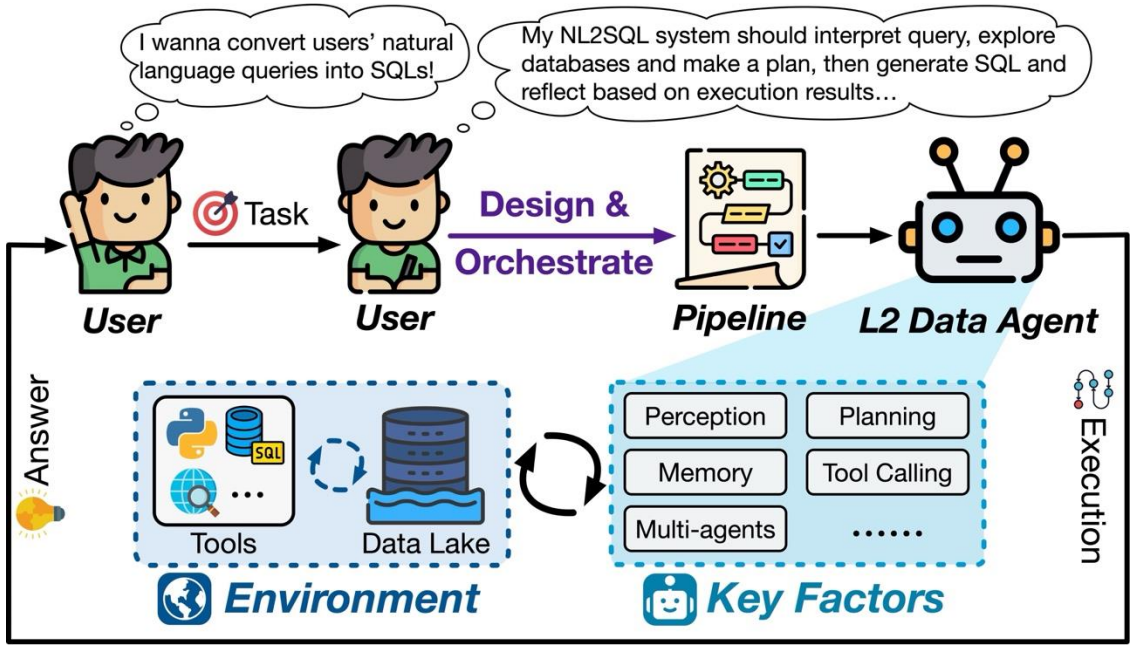
## Progress: Efficiency in Routine Tasks

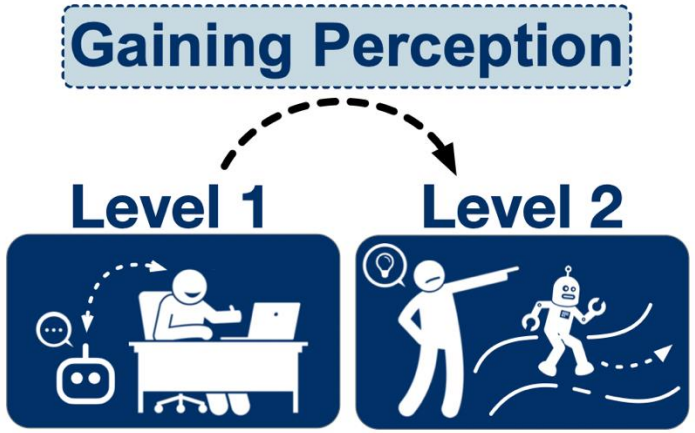
- **Query-Responsive Assistance:** interpret and respond to user queries on demand
- **Efficiency Boost:** improve efficiency by offloading trivial and routine operations, such as unit conversions or standard preprocessing code generation.
- **Lowering Barriers:** lower comprehension barriers for non-technique or novice users

## Limitations of L1 Data Agents

- **Stateless Nature:** operate in a “prompt-response” paradigm without maintaining state over time.
- **Lack of Perception:** unable to perceive or interact with the external environment (databases, APIs) autonomously, preventing a closed-loop refinement and optimization
- **Human Dependency:** The human users still manually execute, integrate, and verify outputs. Data agents cannot perform end-to-end procedures, limiting autonomy to atomic, static subtasks.

# Definition for L2 Data Agents (Partial Autonomy)



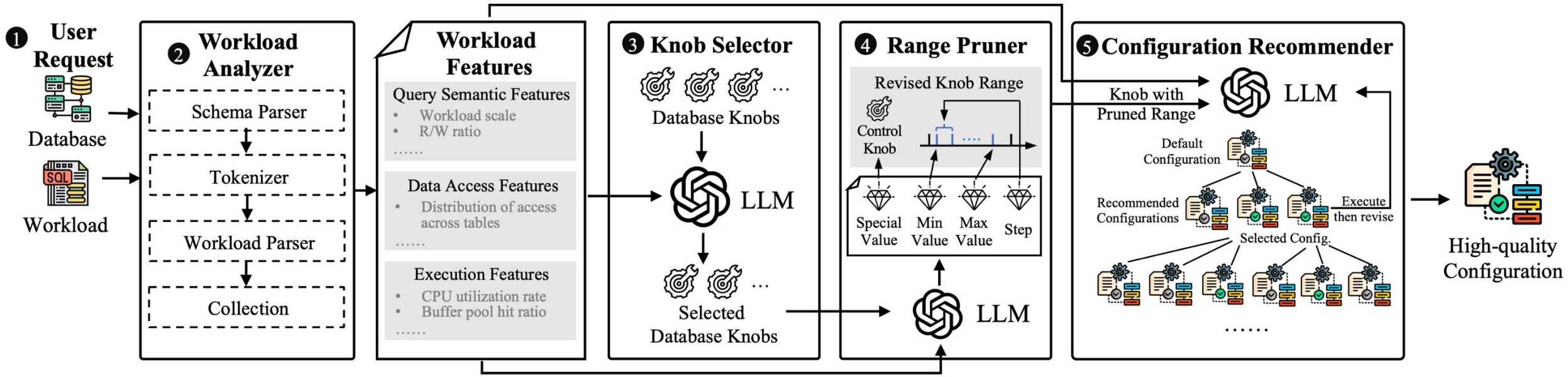




- Data agents can perceive and interact with the environment (e.g., data lakes, DBMS, code interpreters, APIs, etc.), enabling partial autonomy to perform task-specific procedures independently.
- Data agents operate within human-orchestrated pipelines
- The data agent $A$ gains environmental perception and interaction capabilities $(D, E)$, capable of handling specific data-related tasks by executing $(\epsilon_A)$ pipeline $P$ orchestrated by human $H$:

$$H: \pi_H(T, D, E) \rightarrow P. \qquad A: \epsilon_A(P, D, E, M) \rightarrow O$$

# Config Tuning: AgentTune (SIGMOD 2025)

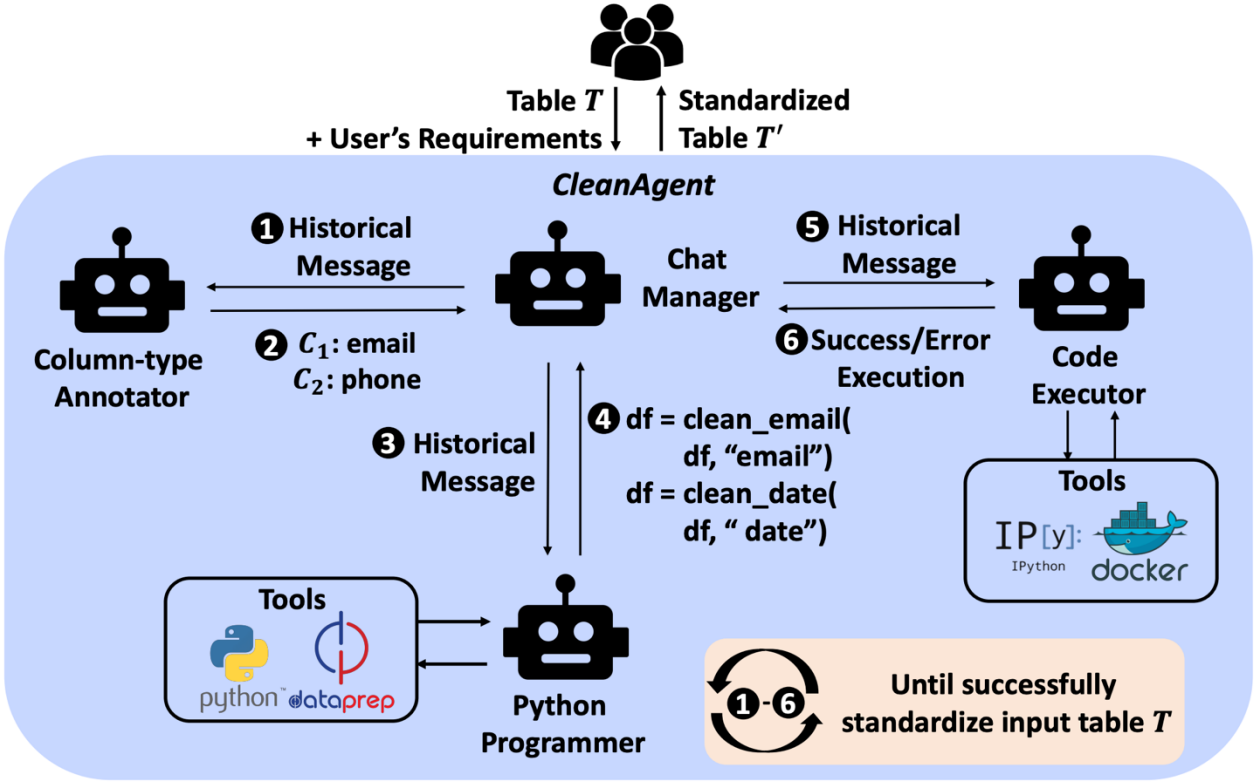
- Decompose the tuning process into specialized agents: Workload Analyzer, Knob Selector, Range Pruner, and Configuration Recommender
- Refine the configuration based on DBMS feedback (performance and execution features) and generate new candidate configurations for a beam search strategy




Li Y, Li H, Zhang J, et al. AgentTune: An Agent-Based Large Language Model Framework for Database Knob Tuning[J]. Proceedings of the ACM on Management of Data, 2025

# Data Cleaning: CleanAgent (VLDB Workshop 2025)

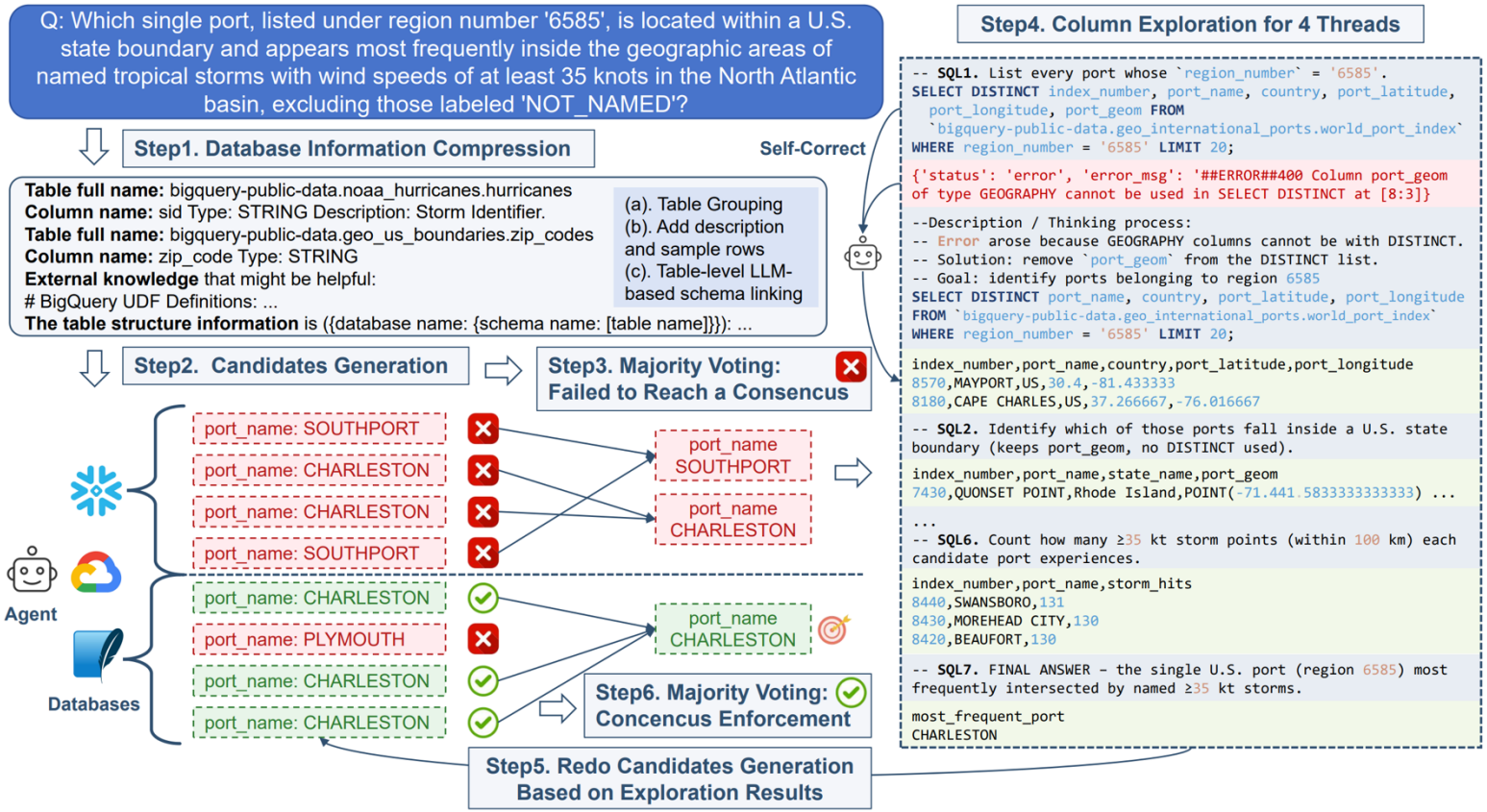
- CleanAgent can interact with raw data and execution environment (like Python or Docker), receiving feedback to refine generated code during data standardization
- Introduce a memory module to maintain the historical conversation context




Qi D, Miao Z, Wang J. Cleanagent: Automating data standardization with llm-based agents[C]. VLDB 2025 Workshop: Data Driven AI.

# Structured Data Analysis: ReFoRCE (arXiv 2025)

- Actively explore and interact with databases through an iterative column exploration mechanism, guided by execution feedback
- Enable progressive self-correction and self-improvement during NL2SQL translation

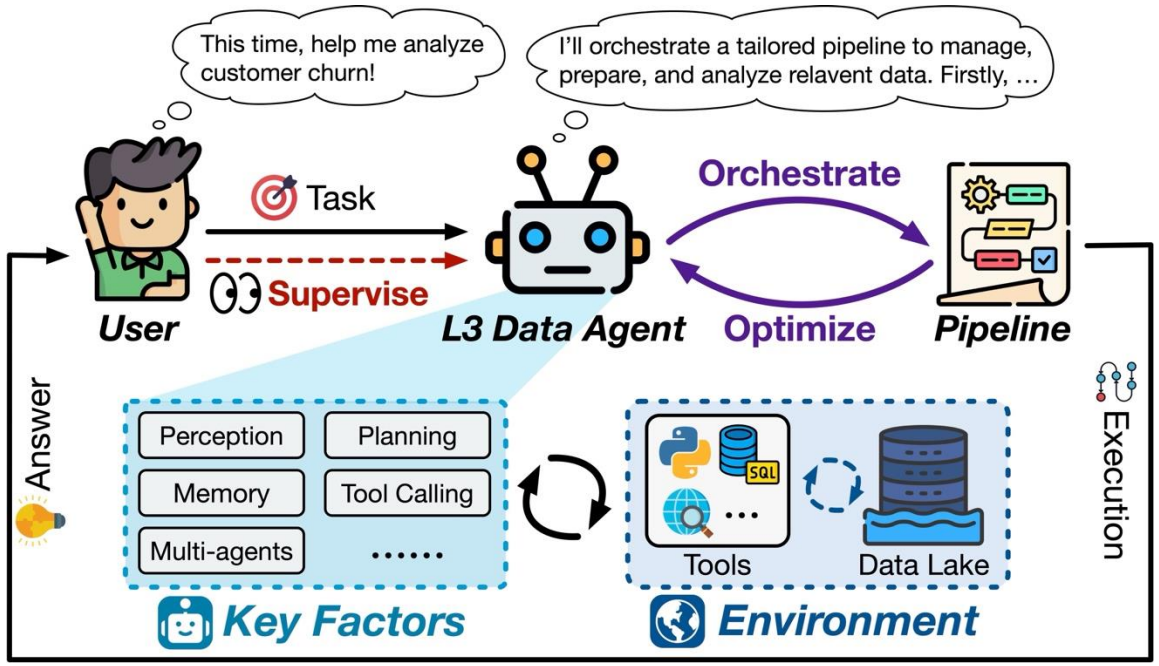


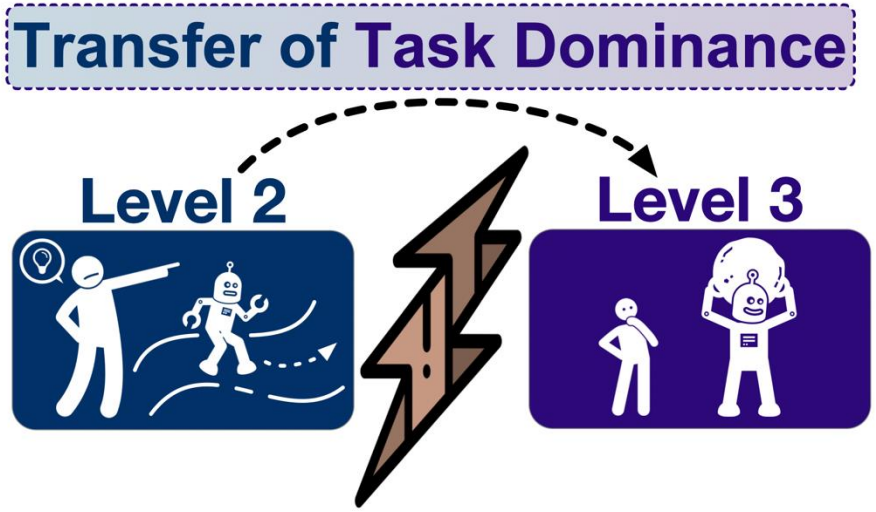

Deng M et al. ReFoRCE: A Text-to-SQL Agent with Self-Refinement, Consensus Enforcement, and Column Exploration[J]. arXiv, 2025.

## The Glass Ceiling of L2 Data Agents (Limitations)

- **Progress — Perception & Interaction:**
  - L2 data agents can connect to real-world systems, autonomously executing specific procedures and optimizing based on environmental feedback
- **Dependence on Human-Designed Pipelines:**
  - L2 data agents comply with pre-established pipelines orchestrated by humans, lacking the ability to independently orchestrate task-tailored new pipelines
  - L2 data agents operate within human-crafted agentic modules, architectures, and collaboration mechanisms
- **Task-Specific Rigidity:**
  - Systems are closely tied to specific tasks/domains (e.g., modules specialized only for NL2SQL)
  - Lack of versatility and generalizability to handle diverse and comprehensive tasks that potentially span the full data lifecycle in real-world scenarios

# Definition for L3 Data Agents (Conditional Autonomy)



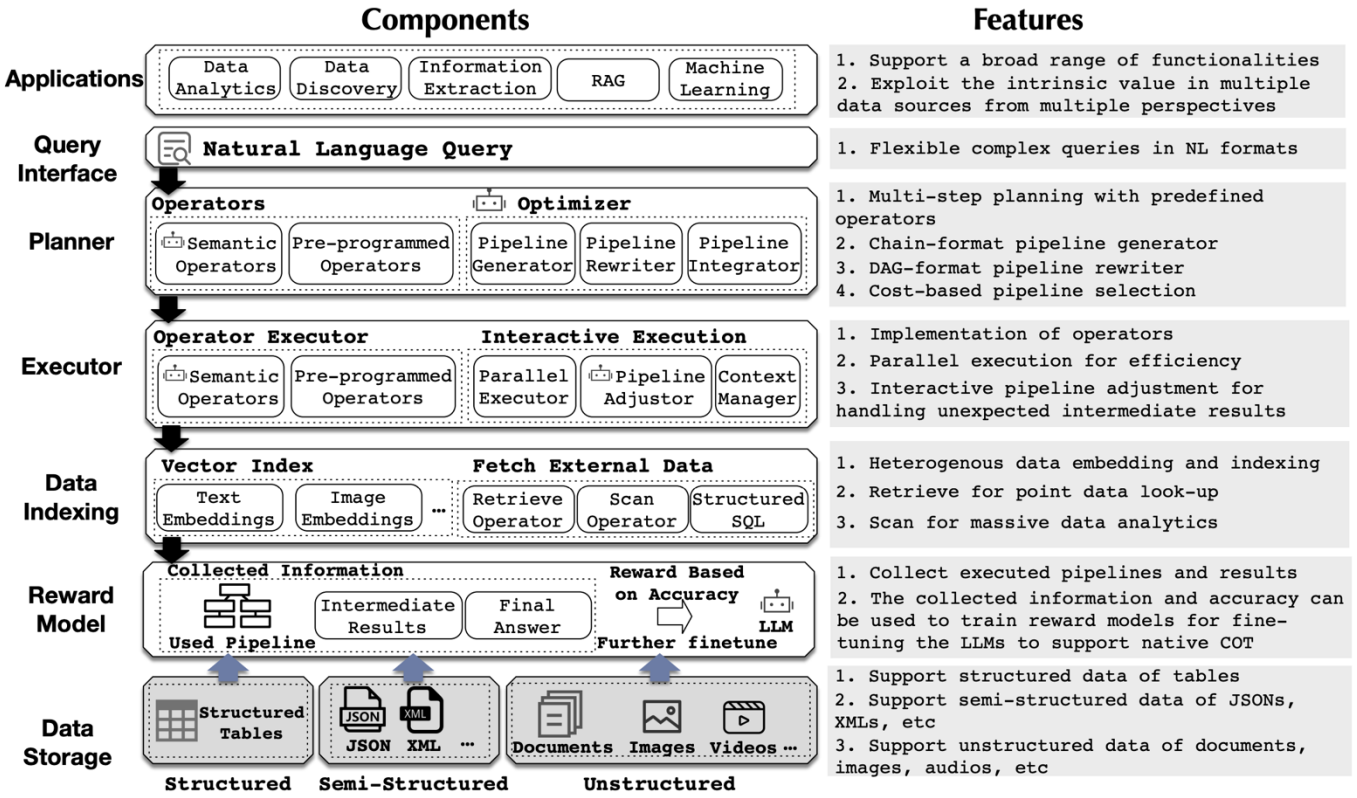

- Data agents autonomously orchestrate and optimize pipelines rather than following human-defined ones; managing diverse and comprehensive tasks potentially spanning the entire data lifecycle, rather than isolated and task-specific procedures
- Critical Leap: data agents assume task dominance from L3, while humans oversee the process.
- Formally, the data agent $A$ autonomously manages the entire pipeline from orchestration $\pi_A$ to execution $\epsilon_A$, tackling versatile and comprehensive data-related tasks $T$ under human $H$ supervision:

$$A: \pi_A(T, D, E, M) \rightarrow P;\ \epsilon_A(P, D, E, M) \rightarrow O. \quad H: Supervise(\pi_A, \epsilon_A)$$

# L3: Striving for Autonomous Data Agents (from Academia)

## Proto-L3: AOP (CIDR 2025)

- **Orchestrate** pipelines with predefined semantic operators for data preparation and semantic analytics across heterogeneous data

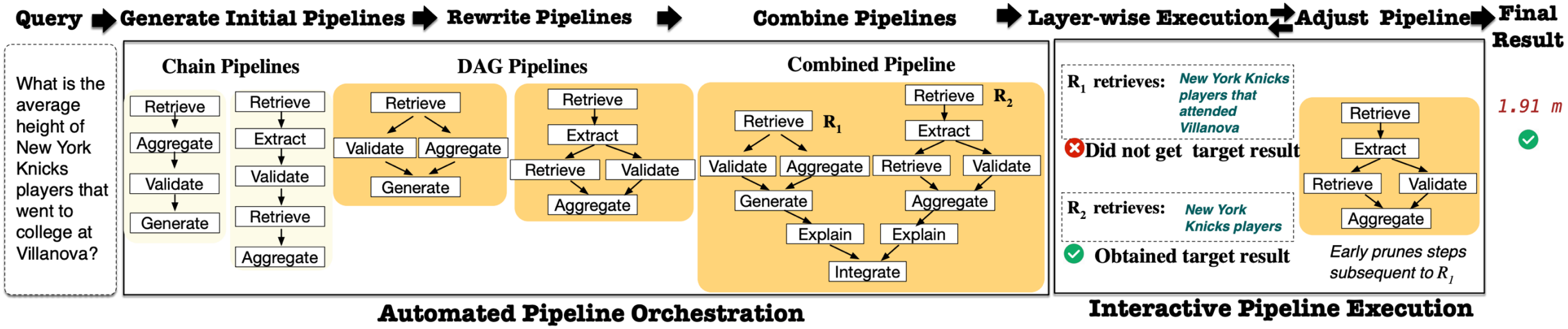



Wang J, Li G. AOP: Automated and Interactive LLM Pipeline Orchestration for Answering Complex Queries[C]. CIDR, 2025.

# L3: Striving for Autonomous Data Agents (from Academia)

## Proto-L3: AOP (CIDR 2025)

- **Orchestrate** pipelines with predefined semantic operators for data preparation and semantic analytics across heterogeneous data
- **Optimize** the initial chain-format execution pipeline into a Directed Acyclic Graph (DAG) to enable parallel execution of independent operators and enhance efficiency

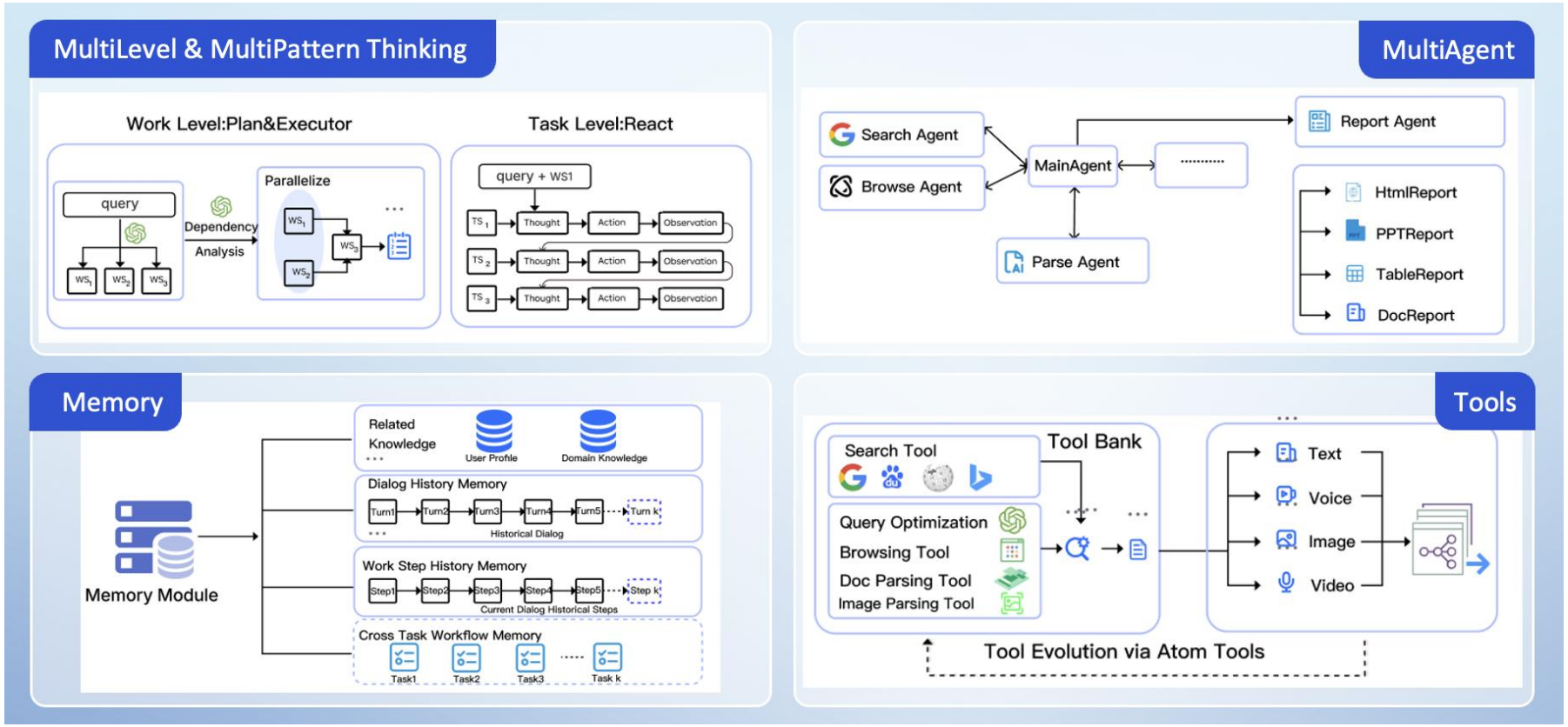



Wang J, Li G. AOP: Automated and Interactive LLM Pipeline Orchestration for Answering Complex Queries[C]. CIDR, 2025.

# L3: Striving for Autonomous Data Agents (from Industry)

## Proto-L3: JoyAgent (JD.com, Inc.)

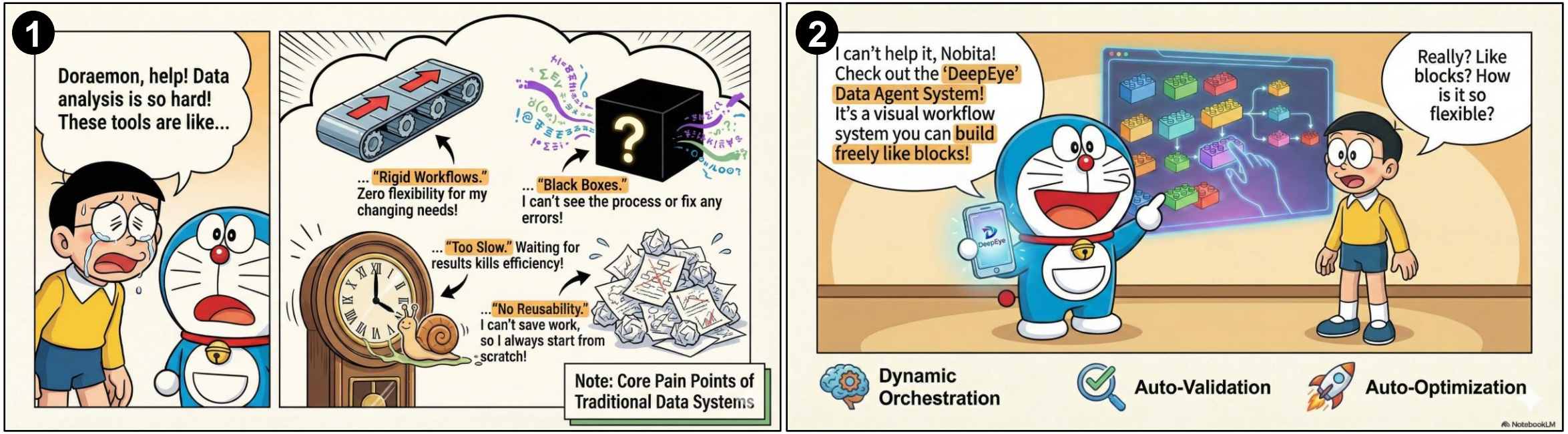
- **Orchestrate** queries into executable DAGs with a multi-level & multi-pattern thinking framework
- **Two Orchestration Modes:** 1) Plan-and-executor; 2) ReAct
- **Tool Evolution:** dynamically creating new tools by recombining atomic ones predefined in tool bank



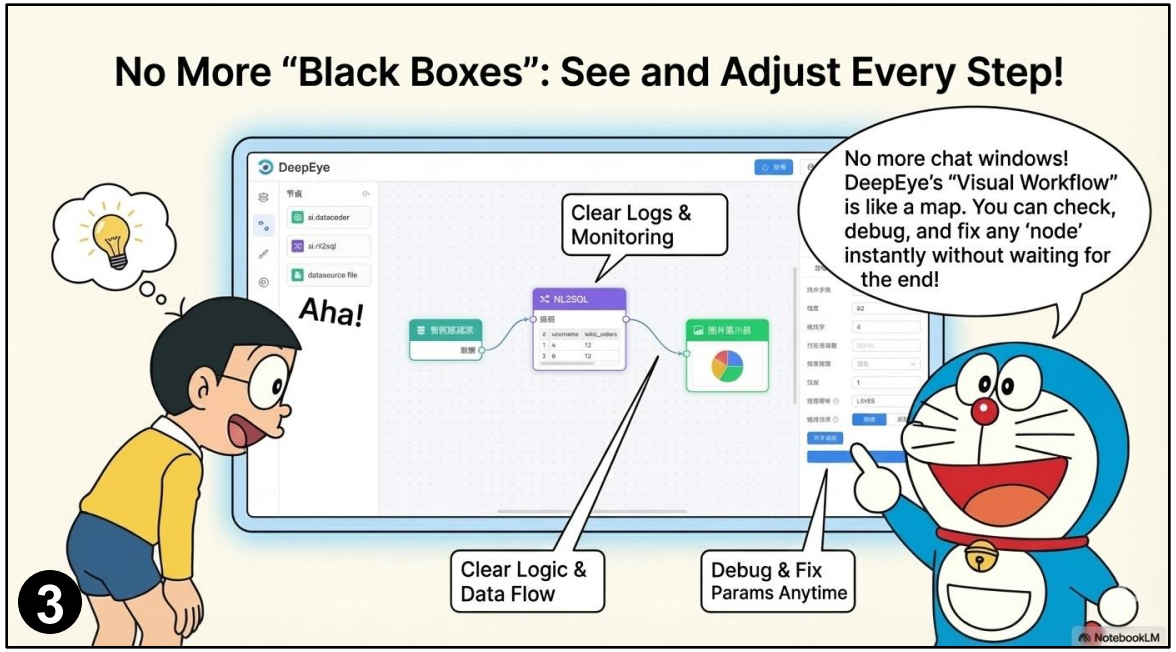

https://github.com/jd-opensource/joyagent-jdgenie/tree/data_agent

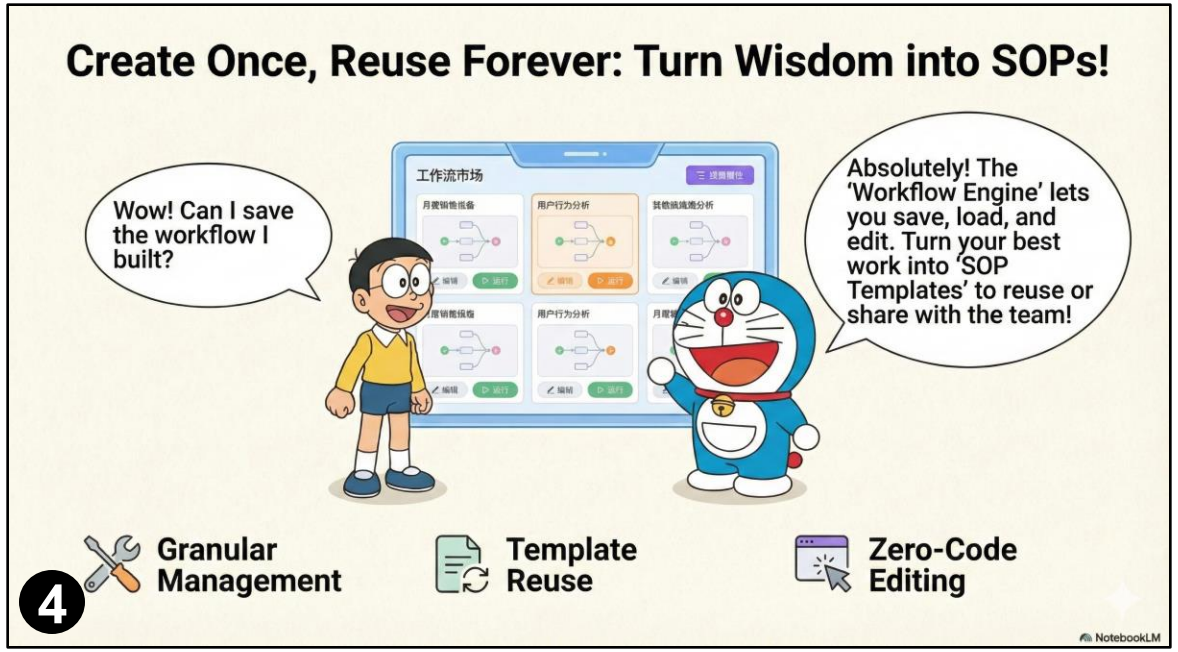
# L3: Striving for Autonomous Data Agents (Our Effort)

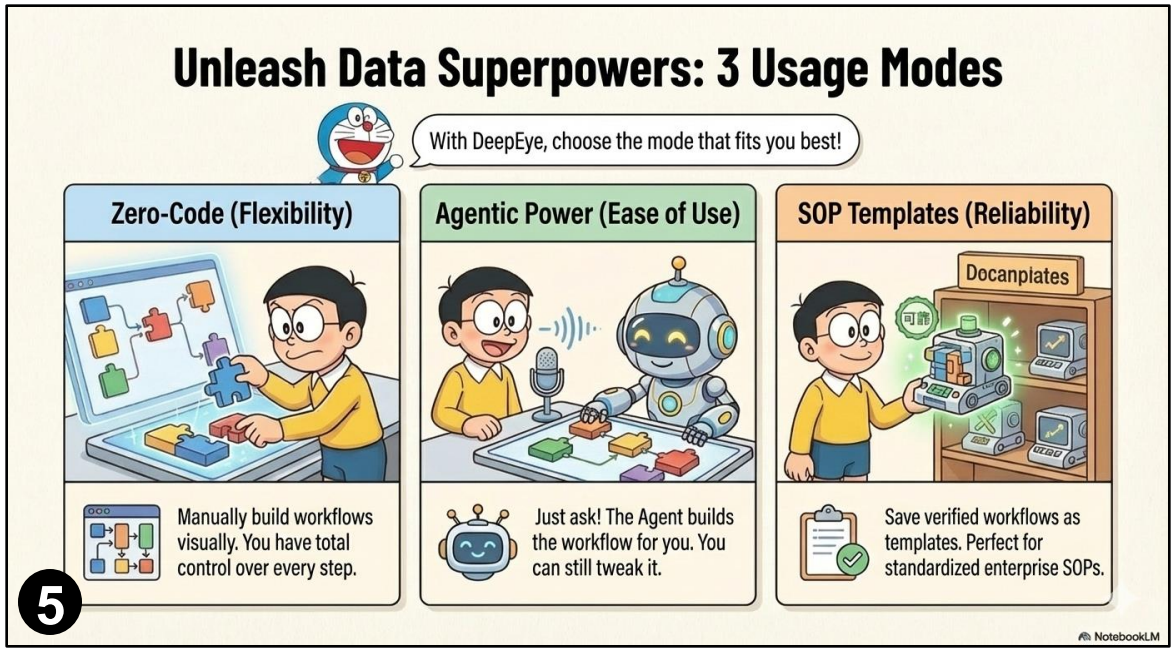
## Our Effort: DeepEye (Stay Tuned!)

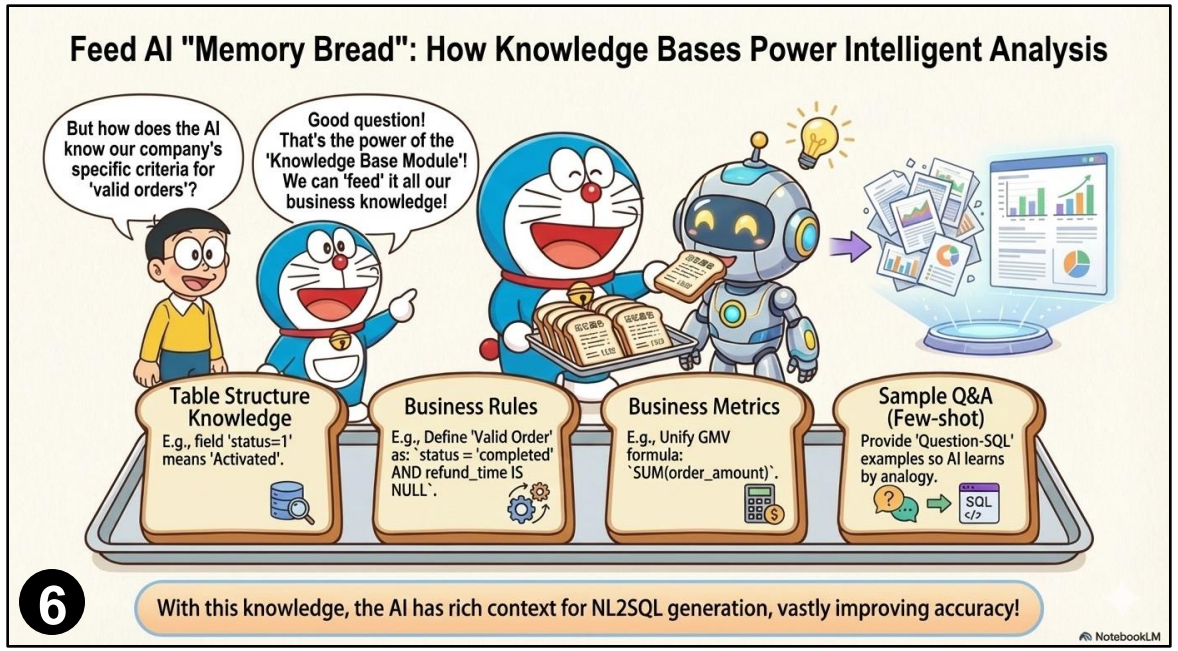


# L3: Striving for Autonomous Data Agents (Our Effort: DeepEye)



## Challenges and Research Opportunities Towards True L3

- **Limited Autonomy in Orchestration:**
  - **Challenge:** Current Proto-L3 systems still rely on predefined operators/tools.
  - **Opportunity:** Skill Discovery [1] and Autonomous Operator Synthesis [2,3]. Autonomously generate, evaluate, and curate new tools/skills dynamically.
- **Incomplete Data Lifecycle Coverage:**
  - **Challenge:** Existing agents focus narrowly (mostly on Analysis, or involve basic preparation), largely neglecting Data Management (tuning, diagnosis) and broader Data Preparation.
  - **Opportunity:** Versatile Generalists. Handle the diverse and comprehensive data-related tasks across the full spectrum in the data lifecycle: Management → Preparation → Analysis.

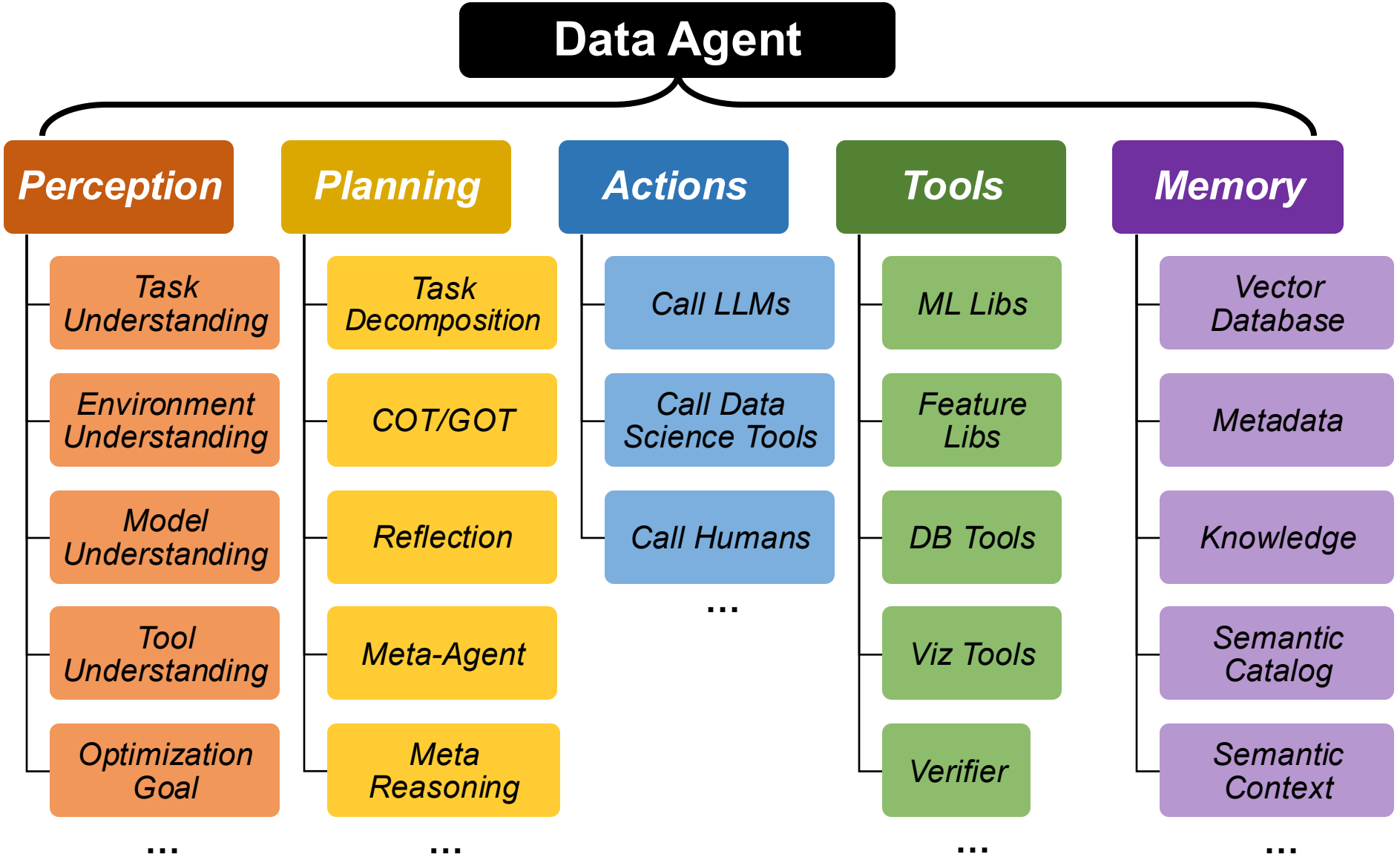
[1] Sun Z, Wang J, Zhao X, et al. Data Agent: A Holistic Architecture for Orchestrating Data+ AI Ecosystems
[2] Sun J, Li G, Zhou P, et al. AgenticData: An Agentic Data Analytics System for Heterogeneous Data[J]. arXiv, 2025.
[3] Wang J, Li G, Feng J. iDataLake: An LLM-powered Analytics System on Data Lakes[J]. Data Engineering, 2025

## Challenges and Research Opportunities Towards True L3 (Cont'd)

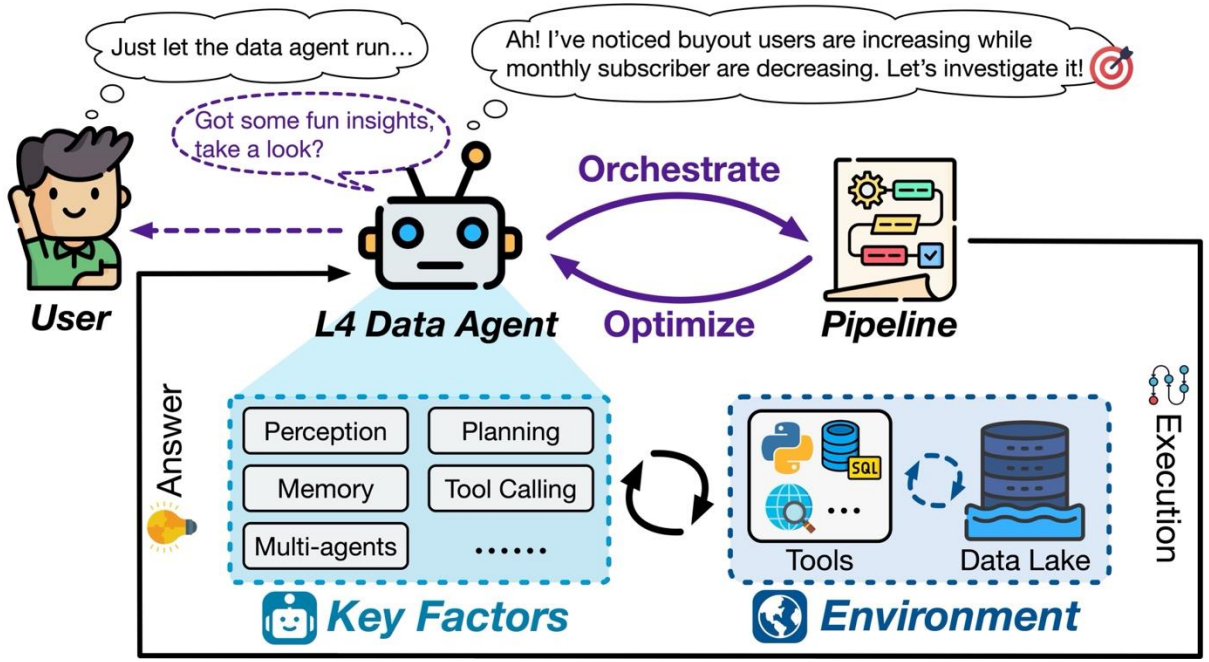
- **Deficiencies in Advanced Reasoning:**
  - **Challenge:** Trapped in tactical fixes and unproductive loops due to a lack of strategic reflection.
  - **Opportunity:** Meta-Reasoning. Incorporating causal reasoning, meta-reasoning for cross-process optimization, and sophisticated memory architectures for abstract strategic knowledge.
- **Adaptation to Dynamic Environments:**
  - **Challenge:** Current evaluations use static data, ignoring real-world data drift.
  - **Opportunity:** Self-Evolution and Dynamic Benchmarking.
    - Enable data agents to adapt to evolving data environments without human intervention.
    - Establish and simulate a dynamic environment to rigorously evaluate and benchmark data agents' robustness under changing conditions

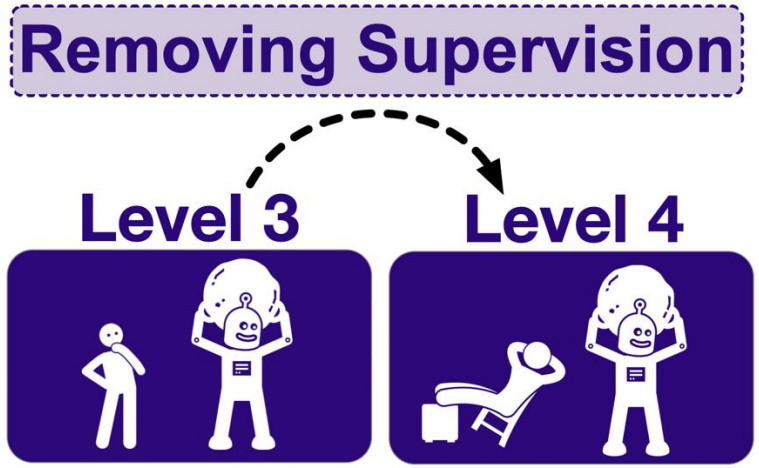
# L3: Striving for Autonomous Data Agents

Data agents need fundamental breakthroughs to achieve L3 autonomy.



# Definition for L4 Data Agents (High Autonomy)



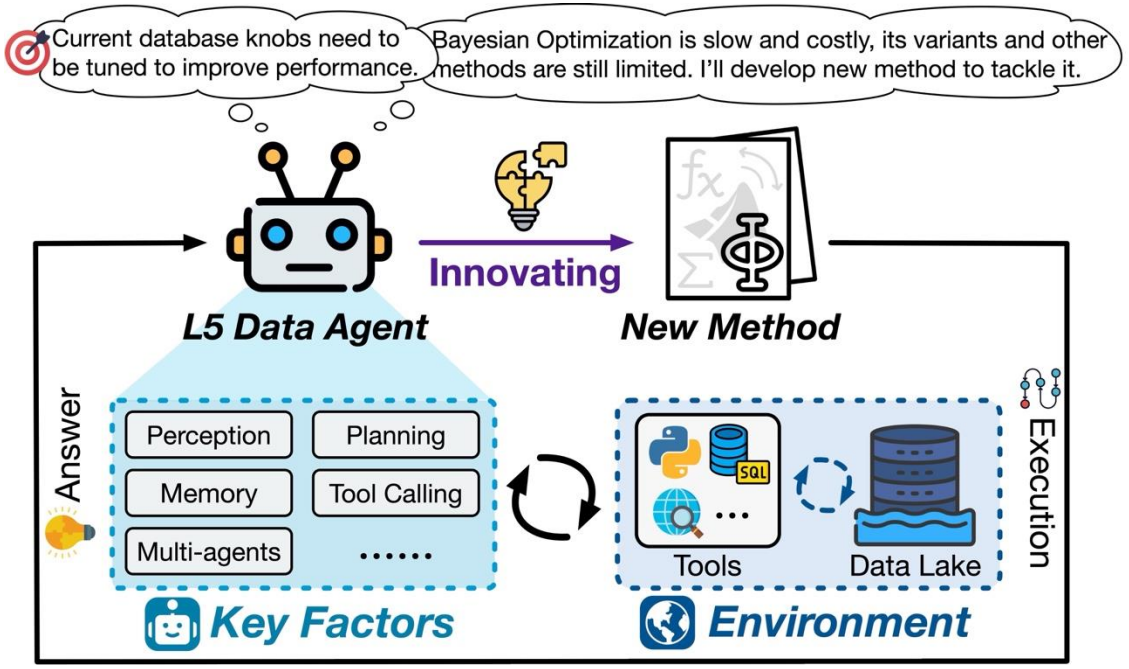

- Data agents autonomously monitor and explore data lakes to proactively identify valuable and emerging tasks, rather than simply responding to given goals/instructions.
- Data agents present high reliability, no supervision is needed, and humans just receive the output.
- Formally, the data agent $A$ takes full initiative, not only orchestrates $\pi_A$ and executes $\epsilon_A$ pipeline $P$ but also autonomously discovers task $T'$ to begin with:

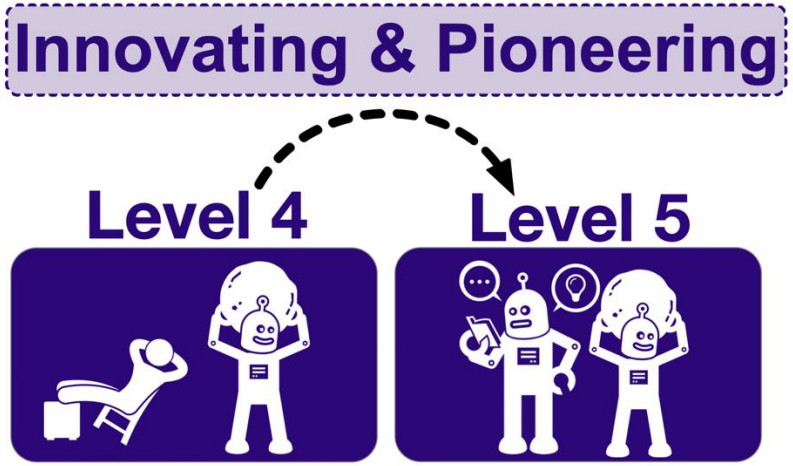
$$A: Discover_A(D, E, M) \rightarrow T';\ \pi_A(T', D, E, M) \rightarrow P;\ \epsilon_A(P, D, E, M) \rightarrow O. \quad H: Receive(O)$$

## Research Directions Towards L4

- **Autonomous Problem Discovery:**
  - Move beyond execution to critical evaluation. Identifying anomalies, gaps, or emerging tasks without explicit task instruction
  - Research Direction: Developing *Task-Oriented Awareness* and *Intrinsic Curiosity*.
- **Trustworthy Self-Governance:**
  - Operate as reliable generalists that orchestrate robust pipelines tailored for self-discovered tasks, and self-manage resources, security, and accuracy without human oversight.
  - Research Direction: Robust effectiveness, efficiency, and safety guarantees
- **Long-Horizon & Holistic Planning:**
  - Make strategic trade-offs (e.g., balancing immediate cleaning costs vs. long-term analytical accuracy).
  - Research Direction: Capabilities for long-term planning and strategic decision-making, beyond local optimizations

# Definition for L5 Data Agents (Full Autonomy)




- Beyond merely applying existing methods, Data agents actively create new knowledge by identifying when conventional approaches are insufficient and innovating novel solutions.
- Formally, the data agent $A$ not only autonomously identifies the promising task $T'$ but also invents a new method $\Phi$ (e.g., a new theory, algorithm, or paradigm) to address it, while human $H$ disengages:

$$A: Discover_A(D, E, M) \rightarrow T';\ \ Innovate_A(T', D, E, M) \rightarrow \Phi;\ \ \Phi(T', D, E, M) \rightarrow O. \quad H: \emptyset$$

# Conclusion

- **Novel Hierarchical Taxonomy for Data Agents**
  - Establishing the first systematic hierarchical taxonomy to resolve terminological ambiguity
- **Structured Lifecycle Review**
  - A structured review of data agents tracing autonomy progression in data-related tasks, mapping the state-of-the-art, and identifying technical challenges and research gaps.
- **Analysis of Evolutionary Leaps**
  - Identifies limitations of current L2 data agents and highlights the critical L2-to-L3 transition as the key research frontier.
- **Forward-Looking Roadmap and Vision**
  - Detailing promising future directions and visions toward proactive and ultimately generative, fully autonomous data agents.

香港科技大学(广州)
THE HONG KONG
UNIVERSITY OF SCIENCE AND
TECHNOLOGY (GUANGZHOU)

# Thank you!

Repo: https://github.com/HKUSTDial/awesome-data-agents
Paper: https://arxiv.org/pdf/2510.23587

Check Paper List!

**Any Questions?**