AI-Ready Open Data Ecosystem

Wenwey Hseush, Shou-Chung Wang, Yong-Yueh Lee and Anthony Ma
BigObject Private Limited, 11 Biopolis Way, #09-03 Helios, Singapore

Keywords: Agentic AI, Retrieval-Augmented Generation, Information Retrieval, AI-Ready Data, AI-Readiness, Data Quality, Data Provision, Analyticity, Large Language Model.

Abstract: AI-ready data sharing plays a pivotal role in the data economy, where data consumption and value creation by AI agents become the undetected new norm in our daily lives. This paper aims to establish an AI-ready, open-ended data infrastructure that spans the Internet to serve live and ubiquitous data to AI agents readily without the repetitive data acquisition and preparation by the agent developers. To achieve this, we propose a standardized framework for evaluating AI-ready data provisioning services, defining the criteria of a data ecosystem and its provisioned data to meet the real-world needs of AI agents.

1 INTRODUCTION

Recent technological breakthroughs culminated by ChatGPT have worked wonders for the world. Behind the wondrous intelligence is a model of collective wisdom that is cultivated from accessible data throughout human civilization. As it quietly integrates into our daily lives, an unprecedented paradigm shift is underway. The new wave of AI-powered applications, commonly known as agentic AI or simply agents, are reshaping the way we live, work and learn. At the core of agentic AI is a grounding process, which enables large language models (LLMs) to adapt to real-world situations to make actionable decisions by accessing and utilizing timely, relevant and contextual data. Note that this grounding data used in agentic AI is distinct from the training data used to develop LLMs.

It is our view that the growing popularity of data-driven AI agents, which is supplanting the existing procedural programming applications, will dominate the expanding landscape of data consumptions going forward. This tidal wave has already created challenges regarding the efficacy of the existing data provisioning mechanism.

1.1 Closed Data Paradigm

While today's data is diverse, live and ubiquitous, mainstream data analytics and data processing remain in the last-century paradigm, where data are collected and prepared by data users, causing unbridled data replications that are time and resource consuming for application development. The user-driven, compute-centric data processing practice referred here as *closed data paradigm* has also raised the attention from global communities concerning the sustainability of our environment.

As AI continues to gain traction around the world, phenomenal insights uncovered through synergies of diverse data across multiple locations have drawn attention worldwide, sparking significant interest and efforts to create data-sharing mechanisms (Kılınç & Küpçü, 2015). However, inherent concerns about security and trust, amplified by widespread data breaches, have made individuals increasingly reluctant to share and use data, imposing serious constraints on data exchange (Zheng et al., 2018). Meanwhile, privacy regulations enacted in recent years, further limited data provisions and complicated information sharing. This explains why the progress of the data economy has been unimpressively slow, and consequently, impeding the advancement of actionable AI.

The closed data paradigm has been deeply rooted in the mind of data workers for decades. "No data, no insights." Is there any other way for sharing data and gaining insights? And why are people so reluctant to share their data? The inertia warrants a new data provisioning solution.

1.2 Open Data Paradigm

Since 2023, we have been experimenting the next-generation data paradigm—a provider-driven, data-centric approach that not only addresses these issues

Hseush, W., Wang, S.-C., Lee, Y.-Y., Ma and A.
AI-Ready Open Data Ecosystem.
DOI: 10.5220/0013647100003967
In *Proceedings of the 14th International Conference on Data Science, Technology and Applications (DATA 2025)*, pages 701-708
ISBN: 978-989-758-758-0; ISSN: 2184-285X
Copyright © 2025 by Paper published under CC license (CC BY-NC-ND 4.0)

but also seeks to tackle the root causes such as data ownership concerns behind data-sharing reluctance.

In this paper, we present an *open data paradigm*, embodied in an open data ecosystem called Aralia. It follows the Web3 principles, featuring a decentralized framework that aligns with the dynamic nature of modern data to serve the development of AI agents. The main property behind the decentralized environment is its ability to extract synergetic insights from multiple sites without downloading or copying datasets, thereby alleviating concerns about data privacy, integrity, etc. Data providers would be open to facilitate secured access to other data users and unlock the opportunities of exploring diverse data, all the while adhering to environmental principles.

An open data ecosystem thrives on its viral interplay between data providers and data consumers, enabling data to attract more data without redundancy. At the core is the sustainability driven by a shared responsibility model, where data providers commit to maintaining high-quality datasets, and consumers engage in responsible data usage. Mutual benefit ensures trust, innovation, and lasting viability.

1.3 AI-Ready Data Provision

This initiative aims to build an AI-ready, data service infrastructure for AI agents to access the live and diverse data worldwide without the acquisition and preparation efforts by agent developers.

The key question that remains is what defines “AI-ready data provision”. An AI agent that relies on a data provisioning service must trust the service’s quality and its governance over time. Besides modelling the Aralia open data ecosystem, we present a criteria framework for AI-ready data provision, covering three facets of AI readiness: service infrastructure, data, and application framework.

In this paper, we focus on the structured data (i.e., tabular datasets), which constitutes a substantial portion of grounding data and serves as a critical foundation for data analysis and AI applications. While our focus is on structured data, it is worth noting that the challenges of unstructured data have been thoroughly addressed in the contexts of the World Wide Web and AI, both of which provide a wealth of resources for AI agents to access.

2 BACKGROUND

We began our research after observing the inertia of the data economy and the paradox it embodies. The paradigm shift ignited by the LLM and agentic AI tsunami further amplified the issues related to data provisioning, which expanded our focus into covering an AI-ready data provision framework.

2.1 AI-Ready Data

Recent research on AI-ready data focuses its role on AI-driven applications, emphasizing attributes such as real-time availability, interoperability, and contextual relevance. Whang et al. (2023) found significant portion of machine learning process in data preparation and studied data validation, cleaning and integration techniques for data-centric AI. Machine learning and AI models that rely on large scale training datasets have also put pressures on data quality issues (Jain et al., 2020).

The U.S. National Science Foundation (NSF) released the report "National AI Research Resource (NAIRR) Pilot Seeks Datasets to Facilitate AI Education and Researcher Skill Development", calling for high-quality AI-ready datasets across various fields to directly support AI-related education, research, and model training.

To address AI-readiness, principles of findability, accessibility, interoperability, and reusability (FAIR) for data and artificial intelligence (AI) are adopted to create AI-ready datasets, especially when some disciplines are subject to restrictive regulations that prevent data fusion and centralized analyses, commonly governed by federal regulations, consortium-specific data usage agreements, and institutional review boards (Chen et al., 2022). These restrictions have spawned the development of federated learning approaches, privacy enhancement technologies (PET), and the use of secure data enclaves. Developing AI models by harnessing disparate data enclaves will only be feasible if datasets adhere to a common set of rules, or FAIR principles.

Meanwhile, a growing emphasis on the importance of ability to include external data in piecing together a broader understanding for analytical goals. Without confidence that data is treated “fairly”, data owners may be unwilling to participate in data sharing (Aaron Gabisch & R. Milne, 2014).

The concept of data ownership has become more important with the prevalence of big data, AI, and privacy laws (Asswad & Marx Gómez, 2021). The continuous breaches from data scams, in addition to unintentional leak incidences, have raised grave concerns, calling for data ownership, stewardship and governance measures (Baker, 2024).

2.2 AI Agent Programming Framework

Traditional procedural programming and the emerging agentic AI RAG (Retrieval-Augmented Generation) framework differ significantly in their principles, execution flow, adaptability and most of all, relationship with data. Procedural programming is a deterministic, structured programming paradigm that follows a rule-driven, instruction-by-instruction, top-down approach to problem-solving using functions and procedures. On the other hand, an agentic AI framework, such as Retrieval-Augmented Generation (RAG) is a non-deterministic (or probabilistic) programming paradigm that integrates real-time data retrieval with LLM, allowing dynamic, context-aware, data-driven reasoning and decision-making experimentation (Singh, 2025).

The transition from procedural programming to agentic AI represents a shift from explicitly programmed logic to context adaptive and retrieval-enhanced intelligence. Developing actionable AI agents requires the access to diverse and contextual data dynamically. Given that data is constantly evolving and emerging the key question for the entire agent industry is, how to dynamically discover and utilize data, especially in the realm of structured data.

3 OPEN DATA ECOSYSTEM MODEL

An open data ecosystem is a decentralized model composed of countless data planets. Each planet, identified by an addressable URL on the Internet, is a data object that

1. encapsulates various datasets lawfully possessed and managed by the planet owner (*i.e.,* data provider or data owner), and
2. provides standard interfaces to interact with an agent application or a user in two ways:
    - Application Programming Interfaces (APIs) for agents (data consumers) to query metadata and explore analytic findings.
    - Graphical User Interface (GUI) for agents and data users to visualize the analytic findings.

To safeguard data ownership, the ecosystem is structured with clear rules that prohibit data downloads, preventing replication or otherwise inadvertent exposure of sensitive information, through data encapsulation with “objects” rather than files, inspired by the notion from the object-oriented programming (OOP) paradigm.

Data consumers and providers engage in a continuous supply-demand cycle, interchangeably, through data planets. Providers curate and maintain read-to-use data within their data planets, while consumers simply explore and query them for the information delivered by providers.

In contrast to the user-driven closed data paradigm, the open data paradigm is provider-driven, in which data is prepared once during the data encapsulation process and is subsequently reusable by multiple consumers via standard interfaces, thereby eliminating repetitive data replications.

3.1 Information Discovery Worldwide

Aralia is a data provisioning service focused on structured data from datasets within an open data ecosystem, whereas Google primarily retrieves unstructured data like text and images or semi-structured data like knowledge graph from the web. Both allow users and AI agents to extract analytic information.

While both systems aim to operate on a global scale, Google Search emphasizes broad information retrieval, whereas Aralia is designed to support the following three key mechanisms for structured data access and information discovery:

- Discover (via Aralia Search) – Identify and retrieve previously unexpected yet closely relevant datasets.
- Explore (via Aralia Explore) – Examine, analyse, and extract meaningful findings from a dataset to uncover patterns or trends.
- Synergize (via Aralia Transplore™) – Cross-analyse datasets from multiple data planets, in search for unforeseen synergetic insights. *Transplore* (“*Trans*cend” and “Ex*plore*”) demonstrates Aralia’s interoperability for exploring across data planets.

3.2 Open-Ended, Fair and Sustainable

The Aralia model is open-ended, allowing data providers to freely enter and exit, governed by a “survival of the fittest” dynamic to evolve organically. The ecosystem sustains through the balance of supply and demand. Data providers are responsible for managing their own data, ensuring good data quality to attract data users, whereas data consumers adhere to fair-use principles, accessing data without retaining data beyond permitted use.

An open-ended data ecosystem can be viewed as a data universe composed of numerous non-overlapping stellar systems, each with a central star (*i.e.,* Sun) surrounded by planets held in a cluster as illustrated in Figure 1. Each central star, serving as the planets' authority (PA), provides services for authentication and resource utilization tracking.

PAs play a vital role in ensuring fairness within the ecosystem. They account for every instance of data usage and ensure appropriate compensation to data providers, who are further incentivized to offer higher-quality data and better services.

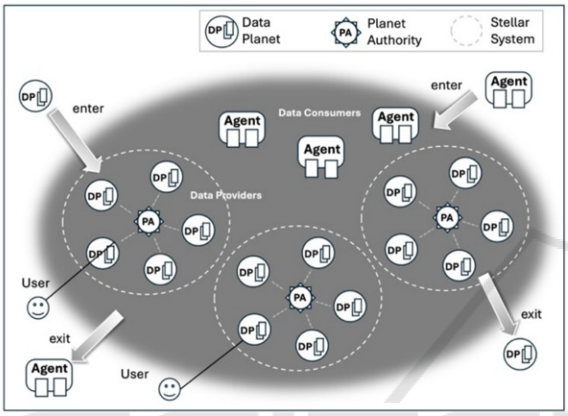

Figure 1: The Aralia Open Data Ecosystem.

As interactions between data providers and consumers grow, the ecosystem fosters a positive feedback loop that continuously attracts new participants and a fair, sustainable market that ensures long-term viability and expansion.

4 AI-READY DATA PROVISION

4.1 Actionable-AI Agents

An AI agent application performs a grounding process by integrating an LLM with real-time, local, and contextual data to accomplish its goals. The LLM applies reasoning, analysing patterns, inferring relationships and drawing conclusions, all based on the data it can find in time. Actionable AI refers to the AI applications capable of constantly making meaningful decisions and taking actions, both reactively and proactively, based on insights adapted from dynamic, real-world conditions. Ultimately, the effectiveness of an actionable-AI application depends on its ability to retrieve accurate and pertinent data from an AI-ready data provisioning service.

4.2 Criteria for AI-Ready Data Provisioning Service

Aralia introduces a framework of criteria for AI-ready data provision, which is designed to assess the quality of a data provisioning service provided to AI agent applications. This approach focuses on addressing data-related issues, leaving system-level concerns such as reliability, scalability, and fault tolerance as assumed defaults without further discussion.

The framework is based on three key pillars, all of which affect the service quality.

1. AI-Ready Service Infrastructure (server aspect):
   - Data Diversity – The variety of data types, sources, and formats available for analysis and processing.
   - Data Analyticity – The ability to extract insights by analysing data.
   - Data Synergy – The enhanced effectiveness and insights gained from the interoperation or interaction of diverse data sources.
   - Equitability – The quality of being fair, impartial, and providing equal opportunities, access or benefits to all stakeholders.
   - Accessibility – The degree of availability for a wide audience to access.
2. AI-Ready Data (data aspect):
   - Timeliness – The availability of data when needed and up-to-date, ensuring it is current and relevant for informed decision making.
   - Completeness – The extent to which all required data is present and fully captured.
   - Validity – The conformity of data to defined standards or schemas.
   - Stability – Consistency and reliability of data content over time.
   - Veracity – Data authenticity and source credibility.
3. AI-Ready Application Framework (client aspect):
   - Industry Standard Compatibility – The ability of a programming framework to align with widely supported industry standards.
   - AI-Ready API – The degree of callable function completeness with AI understandable format and error messages.

Figure 2 shows the AI-Ready Criteria Framework in which, in addition to the abovementioned, a third layer from the left consists of all measures and the fourth layer consists of the mechanisms, each of which corresponds to its measure.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

Based on the criteria framework, a data provisioning service can be evaluated for its readiness to support AI applications and its long-term sustainability. The featured mechanisms used to support the qualities of these three pillars are presented in the following sections.

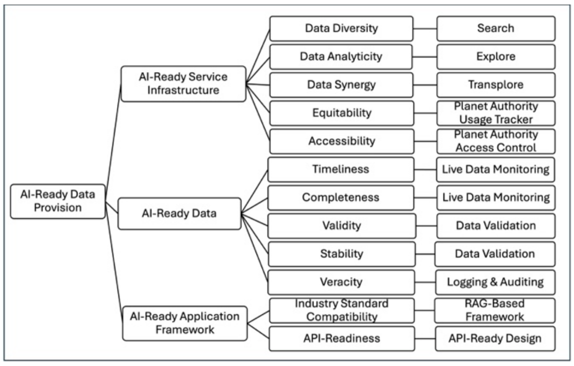

Figure 2: Aralia AI-Ready Criteria Framework.

5 AI-READY SERVICE INFRASTRUCTURE

The goal of an AI-ready service infrastructure is to enable the underlying data platform to deliver accurate grounding data to AI agents. This allows the agents to inquire and process the grounding data and to make informed decisions.

In our case, the underlying data platform that supports the data provisioning service is the Aralia open data ecosystem. In Aralia, we envision an open-ended data universe where data objects are continuously generated and readily accessible anytime, anywhere. It enables AI agents to uncover insightful information by an operation infrastructure with three methods as follows:

- Search: addressing data diversity
- Explore: addressing analyticity
- Transplore™: addressing data synergy

The operation framework enables AI agents to discover relevant datasets (from multiple planets), retrieve analytic information (from one planet), and extract synergetic insights (from multiple planets).

5.1 Search

The strength of the open data ecosystem lies in its vast and unexplored potential, much like the internet. It allows for a multitude of independent data planets to participate, each potentially holding valuable and multifaceted data waiting to be discovered. Driven by human curiosity, the open data paradigm thrives on the belief that countless datasets, both known and unknown, can offer hidden insights waiting to be uncovered. Aralia presumes that the availability of required data is uncertain.

To maintain consistency in accessing the latest relevant data, Aralia adopts a Search method to discover data planets and datasets.

5.2 Explore

The explore method in Aralia enables exploratory data analysis, allowing data users to navigate and reason through tabular data step by step. Instead of relying on traditional SQL queries, Aralia offers an intuitive graphical interface that helps agent developers and general data users visualize analytical findings. Additionally, it provides a robust set of APIs for building applications and AI agents. This policy is based on the following two arguments: (1) no-code/low-code environment for developers based on today's AI data-driven programming paradigm or users with little to no technical skills and (2) no SQL support, which is subject to security compromise, such as SQL injection, which may lead to data integrity and performance concerns.

Exploration is a step-by-step process of reasoning and learning through experimentation. It involves trial and error, and refining solutions until a desired outcome is reached.

Instead of using SQL query, Aralia supports a multivariate approach that breaks down an exploration process into a sequence of operations as follows:

- Unfold a variable (X variable) and check the values (X values)
- Observe the values (Y values) of a second variable (Y variable), each value is corresponding to an X value.
- Examine the X-Y pairs in different ways.
- Filter the current working dataset by setting conditions on X or Y variables, resulting in a new working dataset.

This approach enables agents or users to systematically examine data and discover potential insights. Like a detective, an agent or a user begins a data exploration journey with an analytic question in mind, examining clusters of datasets and deductively interpreting the numbers and evidence to answer the question at hand.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

5.3 Transplore™

At the core of Aralia is the notion of Transplore, meaning to transcend time and space to explore. This neologism signifies an escape journey into uncharted realms beyond the expected, suggesting a mindset of discovery that uncovers extraordinary insights.

An escape journey is initiated during data planet exploration when an agent wants to cross-analyse a related dataset on another data planet. The goal of this cross-analysis is to gain insights on a set of targets of interest (TOI) – the entities the agent is focused on. The relatedness of the two datasets from each planet is established by aligning two variables, one from each dataset that share the same meaning, format, and scale. The TOI comprises of the distinct values of the related variable that underlie the current analysis with which the user is involved. Transplore incorporates a specialized geospatial variable called "Admin Division" from the OpenStreetMap project, enabling cross-analysis of geospatially related datasets.

5.4 Planet Authority – Usage Tracking and Access Control

Aralia employs resource usage tracking and access control to ensure fairness among participants in the open data ecosystem, which serves as one of the most critical factors for its sustainability. The functions are described as follows:

- PA Resource Usage Tracking: addressing equitability (fairness)
- PA Access Control: addressing accessibility

By implementing these measures and mechanisms, the system prevents unauthorized usage, safeguards sensitive data, and promotes a balanced and self-sustaining open data ecosystem.

6 AI-READY DATA

The goal of AI-ready data is to ensure the quality of both data content, and the processes involved in its collection. It encompasses liveness, completeness, validity, stability, and veracity. The data planet owners are responsible for upholding these quality standards and conscious about continuous improvement to ensure a reliable and sustainable data service. These mechanisms are designed to address the issues:

- Live Data Monitoring: addressing timeliness and completeness
- Data Validation: addressing validity and stability
- Logging and Auditing: addressing veracity

6.1 Live Data Monitoring

Live data monitoring for data stream inputs is critical to ensuring the timeliness and completeness of the collected data. Timeliness refers to the frequency and recency with which data is collected and stored, ensuring it reflects the most current state of the data environment. It is crucial for applications that require up-to-date information to make real-time decisions. Completeness evaluates whether the data captures all relevant aspects of your domain. For instance, in an air-quality monitoring network, if only half of the sensors report their readings, the gaps in coverage render the dataset unsuitable for reliable analysis or decision-making.

6.2 Data Validation

Data validity is the first measure for AI-ready data. It can be verified by a data validation process for incoming data, which involves systematically checking for errors such as format or syntax issues that can be identified and corrected using automated methods. These errors may include inconsistencies in data structure, missing values, or incorrect data types, which can be caught through predefined rules or algorithms. However, some errors are more complex and require domain-specific knowledge to identify, as they cannot be detected by simple validation checks. These types of errors often relate to the accuracy or relevance of the data within the context of a particular field and can only be properly validated by experts of the subject matter. Ensuring the correctness of such data is the responsibility of the data providers, who must employ the right tools as well as leverage their expertise to validate the data before it is made available for analysis or decision making.

Meanwhile, during the data validation process, we also examine the properties of data stability. Data stability for an incoming stream refers to the consistency and reliability of data as it flows into a system. Stable data streams exhibit minimal fluctuations in structure and frequency. Factors affecting stability include network reliability, data source consistency, and the handling of missing or delayed data. Implementing buffering, schema validation, and error-handling mechanisms can help maintain stability and prevent disruptions in downstream applications.

6.3 Logging and Auditing

The goal of logging and auditing is to provide a measure for neutrality and transparency for data users. Data origination and modification require meticulous documentation and tracking. Comprehensive records of the data injection process, the encompassing data sources, personnel or application responsible for injections or edits, and timestamps, must be maintained to ensure transparency in data preparation. During data delivery, any undisclosed process that alters the data would render the data provider biased and compromise data integrity and source credibility. To uphold neutrality in the delivery process, data provided to consumers shall be free from any undisclosed filters.

Aralia aims to establish a policy of transparency to support data veracity through providing audit trail recordings to users for future reference.

7 AI-READY APPLICATION FRAMEWORK

The goal of an AI-ready application framework is to ensure that agents and their developers can effectively discover the right data sources and retrieve desired information by following standardized data discovery guidelines. Two approaches are presented:

- API-Ready Design: addressing API readiness
- RAG-based Framework: addressing industry standards compatibility

7.1 API-Ready Design

APIs are critical for the smooth integration of an open-ended data ecosystem within agentic AI workflows. As we stressed previously, ensuring AI-ready API is one of the indications of a high-quality data infrastructure. Standardization through APIs promotes compatibility and productivity, making the ecosystem accessible to a broad spectrum of existing analytical software and emerging AI technologies without extensive adaptation. API readiness thus significantly reduces integration time and costs, enabling faster and widespread adoption.

7.2 RAG-Based Framework

Retrieval-Augmented Generation (RAG) is a technique designed to reduce the likelihood that LLMs generate information that is inconsistent with verifiable real-world facts. The phenomenon of LLMs producing unfaithful or nonsensical text is often termed “hallucination” (Ji et al., 2023) and is a prominent challenge in AI-driven text generation. By pairing an LLM with an external retrieval mechanism that fetches relevant context from a designated knowledge source, RAG has been shown to improve factual consistency in generated content. Upon receiving a question or prompt from the user, instead of the LLM immediately responding, a retriever first searches within a pre-assigned vector database to find contextual data which is most related to the original question. This context is retrieved and combined with the original query and used as a new prompt for the LLM. As a result, this newly generated output should be more satisfactory as it should contain more contextual and relevant data than a simple prompt.

Improving on this method, we extend the original RAG framework with Aralia tools, named OpenRAG in Figure 3. The term “open” preceding “RAG” signifies access to an open-ended data ecosystem. Instead of connecting the LLM to a predefined vector database for unstructured data, we connect it to an open-ended data ecosystem. The premise for this design is that, as the data ecosystem grows organically, so will the knowledge base of the LLM. The advantage of this design is that users of the RAG no longer need to maintain their own databases, which require constant update and maintenance to ensure their quality. Under the OpenRAG framework, agents can access broad spectrum of ready-to-use data via the API calls.

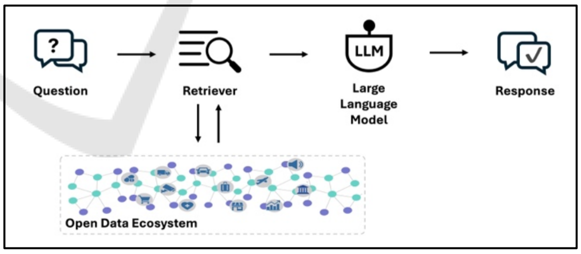

Figure 3: An OpenRAG framework.

7.3 Agent Implementation

AI agents can execute reflectively based on the LLM output. In the probabilistic, data-driven paradigm, AI agent programming greatly reduces the entry barrier for programming.

To illustrate the deployment of agents and their interactions, we designed four AI agents. Acting together, these agents can take a single natural language query from a user, search and select the most suitable datasets from the open data ecosystem, execute and retrieve the raw data, and finally interpret the retrieved data and answer the user's query. Details of the four agents, including their purposes and execution plans, are depicted in Table 1. Each agent is assigned a specific role through its prompt and is organized in a decision graph using LangGraph. All experiments involving AI agents were conducted with the GPT-4o model.

Table 1: Summary of four agents.

Agent	Purpose & Key Steps
Aralia Search Agent	Search for possible datasets and data planets associated to the user's query via Aralia's API.
Analytics Planning Agent	Request LLM to generate an execution plan to answer user's query, provided with meta-information, relevant datasets, interesting variables (x-axis) and attributes (y-axis).
Analytics Execution Agent	Execute the analytic plan by calling the data provisioning APIs to retrieve data insights from the open data ecosystem.
Interpretation Agent	Request LLM to synthesize analytic results into meaningful insights and explain the findings in natural language.

8 CONCLUSION

While everyone agrees with data quality being critical in the modern AI-driven economy, few people pay attention to data provision, and more importantly, AI-ready data provisioning for the viable development of actionable AI. In developing data-responsive applications such as AI agents, their effectiveness relies on the dynamic integration of complementary datasets. To address this need, we introduce Aralia, an AI-ready open data ecosystem, along with a criterion framework for AI readiness. This framework ensures the quality, efficiency, and effectiveness across three key aspects: data, service infrastructure, and application framework, all aimed at enabling AI-ready data provisioning. Based on these criteria, we demonstrate that the Aralia open data ecosystem optimally fulfils the prerequisites for the successful development of AI agents.

REFERENCES

Aaron Gabisch, J., & R. Milne, G. (2014). The impact of compensation on information ownership and privacy control. *Journal of Consumer Marketing*, *31*(1), 13–26. https://doi.org/10.1108/JCM-10-2013-0737

Asswad, J., & Marx Gómez, J. (2021). Data Ownership: A Survey. *Information*, *12*(11), 465. https://doi.org/10.3390/info12110465

Baker, E. (2024, February 17). *Data Ownership and Stewardship in Data Governance*. DataGovernance Platforms.Com. https://www.datagovernanceplatforms.com/data-ownership-stewardship-data-governance/

Chen, Y., Huerta, E. A., Duarte, J., Harris, P., Katz, D. S., Neubauer, M. S., Diaz, D., Mokhtar, F., Kansal, R., Park, S. E., Kindratenko, V. V., Zhao, Z., & Rusack, R. (2022). A FAIR and AI-ready Higgs boson decay dataset. *Scientific Data*, *9*(1), 31. https://doi.org/10.1038/s41597-021-01109-0

Jain, A., Patel, H., Nagalapatti, L., Gupta, N., Mehta, S., Guttula, S., Mujumdar, S., Afzal, S., Sharma Mittal, R., & Munigala, V. (2020). Overview and Importance of Data Quality for Machine Learning Tasks. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3561–3562. https://doi.org/10.1145/3394486.3406477

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Chen, D., Dai, W., Chan, H. S., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, *55*(12), 1–38. https://doi.org/10.1145/3571730

Kılınç, H., & Küpçü, A. (2015). Optimally Efficient Multi-Party Fair Exchange and Fair Secure Multi-Party Computation. In K. Nyberg (Ed.), *Topics in Cryptology — CT-RSA 2015* (Vol. 9048, pp. 330–349). Springer International Publishing. https://doi.org/10.1007/978-3-319-16715-2_18

Singh, A. (2025). *Agentic RAG Systems for Improving Adaptability and Performance in AI-Driven Information Retrieval*. SSRN. https://doi.org/10.2139/ssrn.5188363

Whang, S. E., Roh, Y., Song, H., & Lee, J.-G. (2023). Data collection and quality challenges in deep learning: A data-centric AI perspective. *The VLDB Journal*, *32*(4), 791–813. https://doi.org/10.1007/s00778-022-00775-9

Zheng, X., Cai, Z., & Li, Y. (2018). Data Linkage in Smart Internet of Things Systems: A Consideration from a Privacy Perspective. *IEEE Communications Magazine*, *56*(9), 55–61. https://doi.org/10.1109/MCOM.2018.1701245