# POSITION: WHAT'S THE NEXT FRONTIER FOR DATA-CENTRIC AI? DATA SAVVY AGENTS!

**Nabeel Seedat *♣, Jiashuo Liu *♦, Mihaela van der Schaar♣**
**♣ University of Cambridge ♦ Tsinghua University**

## ABSTRACT

The recent surge in AI agents that autonomously communicate, collaborate with humans and use diverse tools has unlocked promising opportunities in various real-world settings. However, a vital aspect remains underexplored: how agents handle data. Agents cannot achieve scalable autonomy without the ability to dynamically acquire, process, and continually evolve their data ecosystems to navigate complex and changing environments. In this position paper, we argue that data-savvy capabilities should be a top priority in the design of agentic systems to ensure reliable real-world deployment. Specifically, we propose four key capabilities to realize this vision: (1) *Proactive data acquisition*: enabling agents to autonomously gather task-critical knowledge or solicit human input to address data gaps; (2) *Sophisticated data processing*: requiring context-aware and flexible handling of diverse data challenges and inputs; (3) *Interactive test data synthesis*: shifting from static benchmarks to dynamically generated interactive test data for agent evaluation; and (4) *Continual adaptation*: empowering agents to iteratively refine their data and background knowledge to adapt to shifting environments. While current agent research predominantly emphasizes reasoning, we hope this work inspires a broader reflection on the role of data-savvy agents as the next frontier in data-centric AI.

## 1 INTRODUCTION

Imagine a world where AI systems seamlessly collaborate with humans, autonomously gathering and analyzing vast amounts of data, continuously adapting to shifting environments, and providing real-time insights to inform critical decisions. These intelligent agents could proactively assist in areas ranging from healthcare to climate change, enabling smarter policies, personalized education, and faster scientific discoveries. They would act as trusted partners, augmenting human decision-making at every level and responding dynamically to unforeseen challenges.

This vision of AI, capable of self-sustained learning and autonomous action, is no longer purely speculative. Recent developments in large language model (LLM)-based agents have brought us closer to this future, with these systems demonstrating impressive abilities in natural language understanding, problem solving and even tool use. However, despite their remarkable progress, current LLM-based agents are still limited by certain constraints. They mainly operate in controlled environments with predefined data and well-structured tasks, and they rely heavily on static datasets and benchmarks. However, today's AI agents are increasingly expected to operate in open-ended, dynamic environments—whether for scientific discovery, industrial automation, finance, or healthcare. In these scenarios, data presents multiple challenges: it is often *incomplete*, requiring proactive information seeking; *messy and noisy*, demanding sophisticated diagnostic and processing capabilities; *constantly evolving*, necessitating continuous knowledge updates; and *difficult to evaluate* through traditional static benchmarks. This leaves a critical gap: real-world scenarios demand more than predefined knowledge and tasks—they require systems that can actively engage with dynamic, noisy, and evolving data in real-time.

To bridge this gap, we propose a new research direction: the development of data-savvy agents, building on the common ground of agent-based AI (focused on decision-making and automation) and

*NS and JL contributed equally.

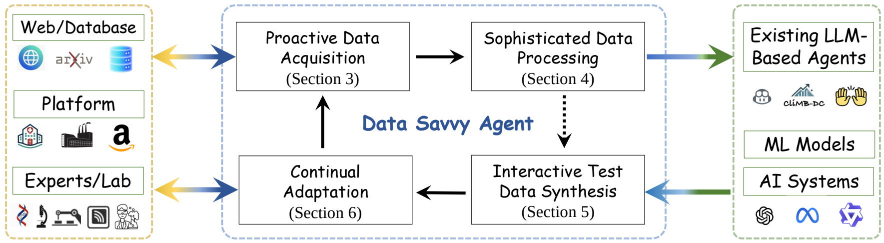


Figure 1: Overall framework of a data-savvy agent, which is built upon four key capabilities: *proactive data acquisition* (see Section 3), *sophisticated data processing* (see Section 4), *interactive test data synthesis* (see Section 5), and *continual adaptation* (see Section 6).

data-centric machine learning (focused on dataset curation in static settings). These data-savvy agents would go beyond merely processing data—they would autonomously acquire, refine, and adapt their knowledge. By advancing autonomous data management, data-savvy agents could enable AI systems to function effectively in complex, ever-changing environments, making them more flexible, resilient, and capable of sustained self-improvement.

In this position paper, we argue that *data-savvy capabilities should be a priority in the design of future agentic systems*. Rather than merely executing tasks with pre-existing data, we highlight four key areas that need focused research: *proactive data acquisition*, *sophisticated data processing*, *interactive test data synthesis*, and *continual adaptiveness*. We believe that by developing agents with these capabilities, we can unlock new opportunities across diverse domains, enabling AI to become a truly transformative force in real-world applications. This paradigm shift unlocks exciting opportunities across multiple research communities: ① **Agent researchers** can extend their focus beyond reasoning and tool use to include dynamic data acquisition, adaptive learning, and real-time knowledge integration, making agents more autonomous and resilient in complex environments; ② **Data-centric ML researchers** can explore new data challenges in open-ended, interactive, and non-stationary environments; ③ **Broader scientific researchers** and **industrial engineers** can leverage data-savvy agents to accelerate discoveries, optimize large-scale industrial systems, and develop AI-driven solutions that continuously learn and evolve alongside real-world complexities.

We hope this position piece stirs debate within the ML research community, challenging the idea that the future of AI will be defined not just by what agents can do—but by what data they can understand, shape, and evolve with — as the path toward truly autonomous agents.

## 2 Data-Savvy Agent

We begin by providing an overview of the role of data-savvy agents. As illustrated in Figure 1, data-savvy agents fill the vacancy between real-world data sources and general LLM-based agents or traditional ML models. Acting as a crucial bridge, they enable the seamless integration of diverse data streams into AI systems.

On one side, data-savvy agents interact with a wide range of data sources, from easily accessible public web databases and open repositories to platform-based data that requires specialized devices or infrastructure—such as hospital records, industrial systems or smart devices. They also handle high-value experimental data, which is hard to obtain, including research findings from laboratories and proprietary curated datasets. On the other side, data-savvy agents engage with AI systems by providing well-curated, high-quality real-time data and performing interactive auto-evaluation, which facilitates continuous adaptation.

To enable this, we posit that data-savvy agents must be equipped with four key capabilities. First, they need *proactive data acquisition*, which involves handling raw, messy, and dynamic data from various sources to gather application-specific data or knowledge. Second, they require *sophisticated data processing*, enabling them to manage and curate diverse data types in a context-aware manner. Third, they must have the ability to conduct *interactive test data synthesis* as well as auto-evaluation, where dynamically generated interaction data helps reliably evaluate agent performance. Finally, data-savvy agents should be capable of *continual adaptation*, which entails iteratively refining data or acquisition strategies to adjust to shifting environments, thus enhancing model performance over time. We detail these capabilities next. Additionally, we provide a structured breakdown of the key research directions and their levels of difficulty in Appendix A.

By integrating these capabilities, data-savvy agents go beyond data-centric AI's focus on improving static datasets or well-structured tasks, to proactive and adaptive data engagement, enabling AI agents to operate in dynamic, real-world settings. For more details on this contrast, see Appendix B.

# 3 Capability 1: Proactive Data Acquisition

*Proactive data acquisition* denotes an agent's autonomous ability to systematically collect relevant data through:

- Automated data acquisition from publicly accessible web sources;
- Targeted data extraction & aggregation from diverse unstructured databases;
- Strategic interactions & experimentation with domain experts and platforms.

This capability emphasizes *resource-intensive*, *dynamic*, and *strategically planned* data gathering, distinguishing it from the relatively straightforward and static nature of information retrieval, as seen in retrieval-augmented generation (RAG) approaches (Gao et al., 2023). Unlike retrieval, which operates on predefined and accessible knowledge bases, data acquisition involves navigating unstructured or restricted data sources, adapting to evolving data landscapes, and managing significant logistical and financial constraints. For a more detailed comparison, please refer to Section 3.2.

## 3.1 Why It Matters?

Data is the foundation of AI/ML systems: high-quality data drives insights and reliability, while poor data undermines even advanced models. For example, ImageNet (Deng et al., 2009) transformed deep learning with large-scale labeled images, while industries like recommendation systems, LLM pretraining, and advertising invest heavily in data collection.

In niche fields such as industrial applications, chemistry, and material science, data acquisition presents significant challenges compared to easily accessible web data. Accessing such data often requires application-specific knowledge, specialized facilities, or even human expertise and experimentation. In fact, for most real-world applications, collecting suitable data is frequently the greatest challenge, often outweighing the complexity of algorithm or model development. Let's consider examples across two domains to demonstrate this: *Environmental monitoring* in pollution source tracking exemplifies fragmentation challenges. Models must reconcile satellite imagery with ground sensor networks, while navigating incompatible agency reporting formats (PDFs vs. APIs). The rarity of labeled crisis data, like chemical spills, further compounds the difficulty of training. Alternatively, in the case of *industrial diagnostics* to predict equipment failure, there can be coverage and quality issues. Legacy facilities often have sparse sensor deployment, while existing sensors suffer from calibration drift and inconsistent logging. Safety protocols and production demands create additional blind spots by restricting when and where data can be collected.

In summary, these hurdles highlight that real-world data collection is rarely straightforward and is frequently shaped by the unique complexities of the application, making it an inherently difficult and nuanced task. As a result, there is a strong practical demand for autonomous data acquisition.

## 3.2 Current Progress & Limitations

Given the critical importance of data acquisition, we begin by reviewing the latest advancements in "traditional data acquisition" within the data-centric AI community. We then highlight the gap between current LLM-based agents and the goal of proactive data acquisition, emphasizing how this crucial aspect has been overlooked and why addressing it is essential—even for the agents themselves.

**Traditional Data Acquisition via Active Learning.** Existing data acquisition primarily focuses on *simulated* or *idealized* settings, where data is often assumed to be pre-collected or available in well-structured and standardized formats. For instance, *active learning* (Settles, 2009; 2011; Konyushkova et al., 2017) addresses the problem of iteratively selecting data points from a large (typically unlabeled) data pool for labeling. However, this approach assumes that the ML researcher has full access to the entire unlabeled data pool—an assumption that does not hold in practice, as the full dataset is typically not visible to the data acquirer. While recent works (Chen et al., 2023; He et al., 2024b) on data acquisition relax the need for full access, they still assume the existence of multiple data sources or providers and primarily focus on efficiently selecting among them. However, this is a secondary problem compared to real-world data acquisition, where *the most significant challenge lies in acquiring a comprehensive data pool* in the first place—as illustrated in Section 3.1.

In addition to traditional data acquisition, recent LLM-based agents have explored information retrieval—a related but simpler, preliminary step in proactive data acquisition.

**Retrieval-Augmented Generation for LLMs** Recently, retrieval-augmented generation (RAG) for LLMs has seen rapid development. This approach enhances LLMs by retrieving relevant document chunks from external knowledge bases during inference, fine-tuning, and even pretraining. The knowledge base typically consists of structured databases or well-curated sources like Wikipedia and arXiv (Gao et al., 2023, Table 1), enabling the LLM agent to retrieve the most relevant documents for more accurate and contextually grounded generation. However, data acquisition is *far more complex* than information retrieval, particularly in the following aspects:

- Data Accessibility: Unlike retrieval, where the knowledge base is predefined and easily accessible, data acquisition often involves navigating complex website structures, dynamic data sources, and negotiating access to proprietary or restricted datasets. Simple search is often insufficient, requiring more sophisticated techniques to extract and integrate relevant data.
- Data Quality and Structure: Retrieval assumes well-structured and high-quality data, whereas acquisition frequently deals with raw, unstructured, or incomplete data that requires significant preprocessing.
- Dynamic and Evolving Sources: Retrieval typically operates on static and pre-defined knowledge bases, while acquisition must account for dynamic, real-time data sources that evolve over time.
- Cost and Scalability: Acquiring large-scale datasets often incurs substantial costs and logistical challenges, unlike retrieval, which leverages existing repositories.

These challenges highlight the limitations of traditional active learning, which often overlooks these practical issues. Additionally, we argue that RAG (retrieval-augmented generation) for LLMs also necessitates proactive data acquisition to enrich and broaden the knowledge base—a point we explore in detail in Section 7.1.

### 3.3 Research Directions

Based on the type of data acquisition—such as whether the data can be obtained from the web or relies on instruments, equipment, or human effort—we outline research directions to empower data-savvy agents on proactive data acquisition.

**Direction 1: Application-Specific *Web Data Acquisition*.** For domains such as biology, chemistry, and climate science, data is often messy, exists in diverse formats, and originates from soures with complex structures, highlighting the need for application-specific data acquisition. We emphasize the following key capabilities that must be addressed:

1. **Application-Specific Data Source Discovery**: The ability to autonomously identify and explore relevant data sources from the web, especially those with *application-specific structures* or *domain-specific nuances*. Many specialized websites, such as biological databases or scientific repositories, feature unique architectures, including dynamic content, nested navigation, or specialized query interfaces. A data-savvy agent must leverage application-specific knowledge to interpret these structures effectively and adapt its exploration strategies. This includes handling advanced web elements (e.g., form submissions, API calls) and filtering sources based on relevance and accessibility. Additionally, the discovery process should integrate domain-specific rules and human or system feedback to iteratively refine the identification of valuable sources.
2. **Structured Data Extraction**: The capability to extract data from complex and heterogeneous formats, such as HTML tables, JSON files, or PDF documents. This requires

advanced parsing techniques, optical character recognition (OCR) for scanned documents, and machine learning models to handle diverse layouts.

3. **Adaptive Acquisition Strategy**: The ability to dynamically adjust acquisition strategies based on real-time feedback or constraints encountered during the process. For example, the agent should recognize when a data source becomes inaccessible, requires alternative access methods (e.g., API authentication), or needs additional domain-specific context to proceed. This ensures the agent can maintain robust performance across varying conditions and prioritize high-value data sources while minimizing time and resource consumption.

**Direction 2: *Experimental Data Acquisition* via Human or Platform Interaction.** When dealing with experimental data in scientific domains like biology, chemistry, materials science or industrial applications, it is often not possible to directly obtain the required data from the web. Instead, agents must interact with human scientists or the platforms for data collection, which could involve the agent assisting in designing experiments and suggesting optimal measurement strategies. By engaging in iterative and context-aware dialogue with human experts, the agent can bridge the gap between computational models and experimental practices, ensuring alignment with the specific requirements of the task at hand. This requires the agent being able to *understand domain-specific experimental constraints* and *adapt their guidance based on feedback or evolving experimental outcomes*. Through such human-agent collaboration, experimental data acquisition can become more efficient, targeted, and aligned with real-world challenges.

## 4 CAPABILITY 2: SOPHISTICATED DATA PROCESSING

*Sophisticated data processing* denotes an agent's autonomous ability to handle complex, real-world data using the following capabilities:

- Diagnose and resolve data issues.
- Adapt processing to contextual nuances and domain specific requirements.
- Integrate and appropriately use advanced data-centric tools.
- Reason both autonomously and with humans about data quality challenges.

This capability emphasizes an agent's ability to reason about and handle complex, real-world data beyond standard preprocessing pipelines.

### 4.1 WHY IT MATTERS?

The promise of autonomous AI agents in tasks from drug discovery to market analysis can only be fulfilled insofar as their ability to handle the messy, dynamic reality of real-world data (Zha et al., 2023; Jarrahi et al., 2023; Kumar et al., 2024; Seedat et al., 2023) — realizing this requires the capability for *sophisticated data processing*.

Real-world data is rarely pristine neither is it static — rather it reflects the real-world, hence has errors, biases, ambiguities and context-dependency (Seedat et al., 2023; Renggli et al., 2021; Aroyo et al., 2022; Sambasivan et al., 2021; Jain et al., 2020). Consider the challenge in finance, where an agent analyzing stock market data, to guide actions, must distinguish between *missing values* caused by technical glitches versus deliberate trading halts - a nuance lost on static data processing pipelines that treat all gaps as noise.

The need for data savviness is further complicated in cases of numerous data issues that might exist in open-ended environments like the web. For instance, an agent trying to crawl for information about a restaurant would need nuanced capabilities to handle complex real-world data which might include AI-generated spam, biased user reviews, or outdated information (Roth, 2024; Read, 2024) — this demands reasoning about what is signal and noise.

### 4.2 CURRENT PROGRESS & LIMITATIONS.

**Current Data-Centric Tools.** The data-centric ML community has introduced various tools to tackle common data issues, including imputation methods (Jarrett et al., 2022), data cleaning solutions (Northcutt et al., 2021), and data valuation tools (Seedat et al., 2024; Jiang et al., 2023). However, real-world data issues are far more complex for two key reasons: (1) as noted earlier, many data problems are context-dependent and require domain-specific knowledge; and (2) these issues often co-occur and are inter-correlated, meaning multiple tools must be integrated to effectively address complex scenarios. Moreover, despite such tools, there is a significant gap in terms of integrated, automated pipelines capable of deploying in specialized domains.

Besides, we review two types of AI agents and how they currently handle data. Firstly, we consider those that process data for modeling and predictive purposes. Secondly, those agents that handle data from open-ended tasks.

**Agents processing data for modeling.** Here we consider agent based systems like Data Interpreter (Hong et al., 2024), DS-Agent (Guo et al., 2024), CleanAgent (Qi & Wang, 2024), GPT Code Interpreter, CliMB (Saveliev et al., 2024) etc. The primary challenge is these agents often default to rigid pipelines based on standard data science practices. While appropriate for standard problems — in more complex cases, simply applying standard pipelines without reasoning about the data and context might lead to failures. Consider the following examples where this might fail:

- A healthcare agent imputes missing blood pressure values using population averages, unaware that missingness correlates with patient severity. As shown in Van Buuren et al. (1999) doing so would result in a model underestimating mortality risk and affect the outcomes and decisions based on the survival analysis.
- A financial agent detects ”outliers” in stock prices during a market crash, misclassifying genuine volatility as noise. Portfolio models trained on this sanitized data fail to hedge against extreme risks, amplifying losses.

These limitations stem from agents treating data processing as a *procedural checklist* rather than *contextual reasoning*. Moreover, current agents optimize for workflow completion (e.g. Data interpreter is based on code execution success) over understanding domain-specific challenges. Finally, despite immense progress in tooling and method development from the data-centric ML research community — current agents often do not integrate said state-of-the-art tools.

**Agents processing data for open-ended cases.** AI agents might also be used for more open-ended tasks in contrast to their use for data science, software engineering and ML pipelines. Let’s consider the case of web surfing agents designed to autonomously navigate the web, interact with websites, and process data to complete tasks such as information retrieval, data extraction, and task automation (He et al., 2024a; Koh et al., 2024).

However, a challenge for agents from the data processing perspective is that web content is often unstructured, dynamic, and noisy. Specifically, agents must process visual elements along with messy HTML and JavaScript. Additionally, while agents focus on the automation aspect, a neglected aspect in open-ended tasks like web surfing is that agents must distinguish between signal and noise, such as filtering out irrelevant ads, AI-generated spam, or outdated information. This requires sophisticated data processing capabilities, as well as, context-aware reasoning.

### 4.3 Research Directions

We highlight the following research directions for the ML community necessary to endow data savvy AI agents with the capabilities of sophisticated data processing.

**Direction 1:** To reimagine data processing as *dynamic and context-aware*, agents need the following capabilities.

1. **Diagnose Interdependent Issues**: Agents should be able to detect and resolve composite challenges like missingness combined with temporal leakage. This would require improvements to agent’s causal reasoning Xiong et al. (2024), moreover performing diagnosis in a context-aware manner (Du et al., 2024; Sarker et al., 2022; Dey, 2018).
2. **Orchestrate Adaptively**: Agents should be able to sequence actions or tool usage in an adaptive and context-dependent manner, accounting for domain constraints. Advancements in task decomposition are vital for this (Gabriel et al., 2024; Rasal & Hauer, 2024).

**Direction 2:** Reimagining human-agent collaboration for data processing alignment with human experts, necessitates the following capabilities.

1. **Expert translation**: Agents should have the capability to translate human requirements into executable and verifiable rules via natural language interaction.
2. **Expert alignment**: Agents should have the capability to align with domain experts. In particular, agents should both ascertain when is the opportune time to prompt experts for information/feedback (Feng et al., 2024b), as well as, have the capability to correct based on this feedback. For instance, in finance, traders can recognize valid ”Black Swan” market anomalies (e.g., flash crashes) such that an agent does not misclassify it as an outlier to remove.

**Direction 3:** Integration of data-centric ML research tools. Current AI agents often default to basic tooling when processing data. Hence, it is vital that future data savvy AI agents incorporate tooling advancements from the data-centric ML research community.

## 5 Capability 3: Interactive Test Data Synthesis

*Interactive Test Data Synthesis* refers to an agent’s ability to autonomously generate, refine, and manage evaluation data tailored to specific tasks and domains. This process includes:

- Context-aware generation of synthetic test data tailored to specific applications or tasks;
- Adaptive integration of feedback loops from domain experts to refine and improve data relevance and evaluation accuracy;
- Generation of interactive, human-like test cases for testing scenarios that require nuanced communication or multi-turn dialogues.

This capability emphasizes the critical role of *data-centric* test case generation, blending human insights and synthetic data to continuously refine the evaluation process and ensure that it aligns with the real-world application of AI systems.

### 5.1 Why It Matters?

Effective evaluation is at the heart of enhancing AI system capabilities. Without proper test data, diagnosing weaknesses and identifying opportunities for improvement becomes nearly impossible. In real-world applications, however, evaluation is far from simple. The key challenges to efficient evaluation include:

1. **Scarcity of High-Quality Test Data**: Unlike traditional tasks with readily available data (e.g., image classification), real-world applications often face limited, fragmented, and noisy test data. This scarcity makes generating reliable evaluation data both critical and challenging (as mentioned in Section 3.1).
2. **Complexity of Tasks and Domains**: Agentic systems, being designed for broad use cases, must be evaluated across a wide spectrum of tasks. For instance, evaluations might span diverse domains, such as software engineering, healthcare, and customer service, which require domain-specific test cases and scenarios (Liu et al., 2024; Xu et al., 2024).
3. **Human-in-the-Loop Evaluation**: Modern agentic systems increasingly require human collaboration for evaluation. However, involving humans in testing introduces scalability issues and demands real-time interaction, which complicates large-scale, automated evaluation processes.

These challenges highlight the importance of *automated and adaptive test data synthesis*. The ability to dynamically create and refine test cases ensures that evaluations are both efficient and representative of real-world applications, empowering systems to adapt and improve faster.

### 5.2 Current Progress & Limitations

In addition to the lack of test data in many real-world scenarios (see Section 3.2), we highlight recent challenges in agentic system evaluation, especially in data generation for complex, human-involved tasks.

**Manual Task Generation for Evaluation.** Existing benchmarks for agentic systems, including tasks from diverse fields such as software engineering and gaming, are curated manually (Liu et al., 2024; Park et al., 2023; Xu et al., 2024). However, this manual process is time-consuming and inefficient. For example, curating tasks for a single agentic system can take several months and thousands of person-hours, making it unsustainable for large-scale evaluation. Furthermore, as LLM-based agents (Fourney et al., 2024; Saveliev et al., 2025) are applied to more complex, open-ended tasks, manually designing tasks becomes even more difficult, with the vast variety of potential use cases leading to infinite possibilities for evaluation.

**Human-in-the-Loop Evaluations.** The integration of human feedback has become crucial in evaluating modern agentic systems (Takerngsaksiri et al., 2024; Saveliev et al., 2025). Copilots such as `GitHub Copilot`, `Cursor`, and `CliMB-DC` are designed to assist non-experts with coding tasks (Saveliev et al., 2025). Evaluating these systems, however, requires real-time collaboration with users, which is a complex and time-consuming process. This difficulty is amplified when test users are experts in domains unrelated to coding, such as clinicians or financial professionals.

These challenges underline the need for an automated, scalable approach to generating relevant, interactive test cases that facilitate large-scale, real-time evaluations with minimal manual effort.

### 5.3 Research Directions

To overcome the limitations of manual evaluation and achieve efficient, large-scale evaluation of agentic systems, we propose the following research directions:

1. **Automated and Context-Aware Test Data Generation**: Future research should focus on developing methods for automatically generating and curating diverse, context-aware datasets (e.g., clinical QA datasets, chemistry datasets) or task sets (e.g., real-life planning tasks, customer support interactions, medical diagnosis tasks) tailored to specific application domains. This approach will reduce reliance on time-consuming manual task design and enhance the efficiency of agentic system evaluation. This approach also complements proactive data acquisition strategies in Section 3.3.
2. **Synthetic Test Case Creation for Human-in-the-Loop Testing** As human-in-the-loop testing becomes more critical, it is essential to develop scalable simulation environments that replicate real-world human-agent interactions. These simulations should support multi-turn dialogues and incorporate domain-specific knowledge (e.g., for healthcare or law), enabling detailed evaluations of agentic systems across varied contexts. This could even lead to the creation of evaluation agents—specialized agents that generate and evaluate other agents' performance within specific domains.

By embedding these capabilities into data-savvy agents, we can create more efficient, scalable, and accurate evaluation methods, both for traditional machine learning models and for more advanced, interactive agentic systems.

## 6 Capability 4: Continual Adaptiveness

*Continual Adaptiveness* denotes an agent's autonomous ability to iteratively refine its data, knowledge, and decision-making processes in response to non-stationary environments. This includes the following capabilities:

- Incremental Knowledge Updating — both knowledge bases and data ingestion processes.
- Proactive change detection.
- Retain prior knowledge while integrating new knowledge (plasticity vs stability).

This capability highlights the importance of an agent's ability to adapt and evolve over time, ensuring its ongoing relevance and performance in dynamic environments.

### 6.1 Why It Matters?

Continual adaptiveness—the ability for AI agents to iteratively refine their ingested data and background knowledge bases in response to shifting environments or changes over time—is foundational for agentic systems to achieve real-world relevance. Let us unpack this vital capability:

Real-world environments are non-stationary — constantly shifting or changing over time. Consider the case of the COVID-19 pandemic. An agent operating pre-pandemic vs during the pandemic would need to continually adapt to the latest policy changes, news updates, patient populations, treatment guidelines etc (Bhuyan et al., 2025). Or consider the case of an agent browsing the web — as privacy and data storage regulations change, an agent should be able to continuously and autonomously update its knowledge base, so that actions adhere to regulations. Without this dynamic and continual updating, we risk the case where AI agents either produce sub-par results or do not adhere to the latest guidelines or policies. We note that this requires autonomous continual adaptation to ensure scalability.

### 6.2 CURRENT PROGRESS & LIMITATIONS.

Despite significant progress in agent design, most systems fail to meet the requirements of continual adaptiveness. We highlight two key dimensions pertinent to current AI agents.

Firstly, current agents struggle with knowledge retention when faced with new information, which can lead to catastrophic forgetting when integrating new information (Zheng et al., 2025; Luo et al., 2023; Li et al., 2024; Thakkar et al.). Secondly, even ignoring catastrophic forgetting, current agents lack anticipatory capabilities (Amos-Binks et al., 2023). Specifically, they cannot anticipate environmental changes and cannot proactively update their knowledge bases and data ingestion to account for these changes.

### 6.3 RESEARCH DIRECTIONS

To bridge the gap between current agent capabilities and the demands of dynamic environments, we propose the following research directions for the ML community.

1. **Dynamic Memory Architectures:** A core limitation of current agents is their inability to retain and contextually update knowledge over extended deployments — i.e. continual learning without forgetting. One vital research direction is improving the memory systems beyond approaches like static replay buffers, which fail to balance integrating new information (plasticity) with preserving critical prior knowledge (stability) (Tao et al., 2023). However, scaling such architectures to real-world applications requires innovations to RAG (Tang et al., 2024; Wang et al., 2024b) and *task-aware memory prioritization*.
2. **Proactive Adaptation:** Current agents remain largely reactive (Lu et al., 2024; Bandyopadhyay et al., 2025), updating models only after performance degradation becomes evident. Closing this gap requires frameworks that incentivize monitoring for changes (i.e. via the agents own initiative) (Liu et al., 2023) and reacting to said changes (Corradini et al., 2022). These could involve agents constantly assessing the data for changes or alternatively monitoring proxies (such as news). Beyond simply identifying changes, a key capability is to quantify the value of the proactive update vs potential costs of the update. This is a particularly important capability as in reality, these updates are likely to incur a cost and hence autonomous data savvy agents should be able to quantify the value of the information adaptation in the context of the environmental change.

## 7 REAL-WORLD IMPACTS

In this section, beyond the impacts discussed earlier, we illustrate how data-savvy agents could transform various fields through two concrete future examples. Due to space limits, additional examples on *autonomous policy adaptation*, *personalized and lifelong education*, *precision healthcare* and *supply chains* can be found in Section C.

### 7.1 SELF-EVOLUTION OF LLM-BASED AGENTS

As agents become more autonomous and widely deployed, ensuring continual improvement without human intervention remains a challenge. Traditional AI development relies on periodic retraining with newly collected data, often requiring extensive human oversight. In contrast, a data-savvy agent could *proactively acquire and curate high-quality data, filtering out noise and refining its reasoning* through continuous interaction with users and external knowledge sources. By leveraging *interactive auto-evaluation*, it could enable agents to assess their own performance, identify weaknesses, and iteratively enhance their decision-making—all without direct human involvement.

This paradigm shift would make data-savvy agents more adaptive, resilient, and capable of sustained autonomous deployment across diverse domains. For instance, in *RAG agents*, as suggested by Shao et al. (2024), a data-savvy agent could continuously refine and optimize the underlying database, allowing RAG-based LLMs to evolve autonomously and enhance their retrieval quality over time. In *scientific research agentic assistants*, a data-savvy agent could continuously ingest and synthesize the latest publications, improving its ability to assist researchers in hypothesis generation, experimental design, and knowledge discovery.

### 7.2 "Lab-in-the-loop" for Scientific Research

Scientific discovery is increasingly data-driven, but fields like drug discovery and materials and climate sciences involve too many possible experiments for humans to explore manually. A data-savvy agent could enable a *Lab-in-the-Loop*" paradigm by actively acquiring experimental data, integrating findings, generating hypotheses, designing (or even conducting) experiments and analyzing results. Through interactive auto-evaluation and continual adaptation, it could refine predictions based on experimental outcomes, continuously improving the support for scientists.

In *pharmaceutical research*, an AI agent could autonomously propose and even test molecular compounds, rapidly identifying potential drug candidates. In physics, it could simulate high-energy particle interactions, refining models with real-world collider data. By bridging human scientists and complex experiments, a lab-in-the-loop system could accelerate breakthroughs across disciplines.

Beyond individual fields, these systems could democratize research, enabling smaller institutions to leverage AI-driven discovery. Early efforts (Boiko et al., 2023; Swanson et al., 2024) in chemistry and biology suggest they may evolve into autonomous labs, where agents conduct experiments, analyze results, and formulate new scientific theories.

## 8 Alternative Viewpoints

Two alternative perspectives to the idea of a data-savvy agent warrant discussion.

► The first perspective is that advances in agent reasoning alone could be sufficient for real-world tasks, making specialized data-centric capabilities unnecessary. i.e. if agent reasoning and planning improves, they should naturally have the capability to handle data challenges.

While this holds in controlled settings with well-structured data, real-world environments are dynamic, incomplete, and often biased. Strong reasoning alone cannot compensate for missing or unreliable information. Even advanced frontier models underpinning agents still suffer from hallucinations and inference errors, underscoring the need for proactive data acquisition, validation, and adaptation.

► The second perspective is around technical feasibility. It can be argued that AI agents already struggle with reasoning and tool use—adding autonomous data-savviness would introduce overwhelming technical complexity.

However, this creates a false dichotomy between improving reasoning and developing data-savviness. Many agent failures stem from poor data handling, making it a fundamental necessity rather than just an engineering hurdle. Instead of postponing the challenge, we advocate for a pragmatic approach: focusing on high-impact, constrained domains like scientific research or industrial processes. This allows controlled progress, where advances in reasoning and data capabilities evolve in tandem, each informing the other.

## 9 Conclusion

We believe data-savvy agents represent an essential yet underexplored frontier in AI research — integrating proactive data acquisition, sophisticated processing, interactive evaluation, and continual adaptation. We hope this position piece stirs debate within the ML community to reconsider the foundational role of data in agentic AI. Through the four data-savvy capabilities and research directions proposed, we aim to inspire new research advances towards realizing the vision of data-savvy agents. To facilitate the practical implementation, we provide further actionable research directions in Appendix A, categorized by difficulty level.

## REFERENCES


Adam Amos-Binks, Dustin Dannenhauer, and Leilani H Gilpin. Anticipatory thinking challenges in open worlds: Risk management. *arXiv preprint arXiv:2306.13157*, 2023.

Lora Aroyo, Matthew Lease, Praveen Paritosh, and Mike Schaekermann. Data excellence for ai: why should you care? *Interactions*, 29(2):66–69, 2022.

Saptarashmi Bandyopadhyay, Vikas Bahirwani, Lavisha Aggarwal, Bhanu Guda, Lin Li, and Andrea Colaco. Yeti (yet to intervene) proactive interventions by multimodal ai agents in augmented reality tasks. *arXiv preprint arXiv:2501.09355*, 2025.

Soumitra S Bhuyan, Vidyoth Sateesh, Naya Mukul, Alay Galvankar, Asos Mahmood, Muhammad Nauman, Akash Rai, Kahuwa Bordoloi, Urmi Basu, and Jim Samuel. Generative artificial intelligence use in healthcare: Opportunities for clinical excellence and administrative efficiency. *Journal of Medical Systems*, 49(1):10, 2025.

Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

Tiffany Tianhui Cai, Hongseok Namkoong, and Steve Yadlowsky. Diagnosing model performance under distribution shift. *arXiv preprint arXiv:2303.02011*, 2023.

Lingjiao Chen, Bilge Acun, Newsha Ardalani, Yifan Sun, Feiyang Kang, Hanrui Lyu, Yongchan Kwon, Ruoxi Jia, Carole-Jean Wu, Matei Zaharia, et al. Data acquisition: A new frontier in data-centric ai. *arXiv preprint arXiv:2311.13712*, 2023.

Flavio Corradini, Miichele Loreti, Marco Piangerelli, and Giacomo Rocchetti. Reptile: A proactive real-time deep reinforcement learning self-adaptive framework. *arXiv preprint arXiv:2203.14686*, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Anind K Dey. Context-aware computing. In *Ubiquitous computing fundamentals*, pp. 335–366. Chapman and Hall/CRC, 2018.

Hung Du, Srikanth Thudumu, Rajesh Vasa, and Kon Mouzakis. A survey on context-aware multi-agent systems: Techniques, challenges and future directions. *arXiv preprint arXiv:2402.01968*, 2024.

Sabri Eyuboglu, Maya Varma, Khaled Kamal Saab, Jean-Benoit Delbrouck, Christopher Lee-Messer, Jared Dunnmon, James Zou, and Christopher Re. Domino: Discovering systematic errors with cross-modal embeddings. In *International Conference on Learning Representations*, 2023.

Jean Feng, Harvineet Singh, Fan Xia, Adarsh Subbaswamy, and Alexej Gossmann. A hierarchical decomposition for explaining ml performance discrepancies. *arXiv preprint arXiv:2402.14254*, 2024a.

Xueyang Feng, Zhi-Yuan Chen, Yujia Qin, Yankai Lin, Xu Chen, Zhiyuan Liu, and Ji-Rong Wen. Large language model-based human-agent collaboration for complex task solving. *arXiv preprint arXiv:2402.12914*, 2024b.

Adam Fourney, Gagan Bansal, Hussein Mozannar, Cheng Tan, Eduardo Salinas, Friederike Niedtner, Grace Proebsting, Griffin Bassman, Jack Gerrits, Jacob Alber, et al. Magentic-one: A generalist multi-agent system for solving complex tasks. *arXiv preprint arXiv:2411.04468*, 2024.

Adrian Garret Gabriel, Alaa Alameer Ahmad, and Shankar Kumar Jeyakumar. Advancing agentic systems: Dynamic task decomposition, tool integration and evaluation using novel metrics and dataset. *arXiv preprint arXiv:2410.22457*, 2024.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

Shantanu Ghosh, Rayan Syed, Chenyu Wang, Clare B Poynton, Shyam Visweswaran, and Kayhan Batmanghelich. Ladder: Language driven slice discovery and error rectification. *arXiv preprint arXiv:2408.07832*, 2024.

Siyuan Guo, Cheng Deng, Ying Wen, Hechang Chen, Yi Chang, and Jun Wang. Ds-agent: Automated data science by empowering large language models with case-based reasoning. *arXiv preprint arXiv:2402.17453*, 2024.

Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Yong Dai, Hongming Zhang, Zhenzhong Lan, and Dong Yu. Webvoyager: Building an end-to-end web agent with large multimodal models. *arXiv preprint arXiv:2401.13919*, 2024a.

Yue He, Dongbai Li, Pengfei Tian, Han Yu, Jiashuo Liu, Hao Zou, and Peng Cui. Domain-wise data acquisition to improve performance under distribution shift. In *Forty-first International Conference on Machine Learning*, 2024b.

Sirui Hong, Yizhang Lin, Bang Liu, Bangbang Liu, Binhao Wu, Ceyao Zhang, Chenxing Wei, Danyang Li, Jiaqi Chen, Jiayi Zhang, et al. Data interpreter: An llm agent for data science. *arXiv preprint arXiv:2402.18679*, 2024.

Abhinav Jain, Hima Patel, Lokesh Nagalapatti, Nitin Gupta, Sameep Mehta, Shanmukha Guttula, Shashank Mujumdar, Shazia Afzal, Ruhi Sharma Mittal, and Vitobha Munigala. Overview and importance of data quality for machine learning tasks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 3561–3562, 2020.

Mohammad Hossein Jarrahi, Ali Memariani, and Shion Guha. The principles of data-centric ai. *Communications of the ACM*, 66(8):84–92, 2023.

Daniel Jarrett, Bogdan C Cebere, Tennison Liu, Alicia Curth, and Mihaela van der Schaar. Hyperimpute: Generalized iterative imputation with automatic model selection. In *International Conference on Machine Learning*, pp. 9916–9937. PMLR, 2022.

Kevin Jiang, Weixin Liang, James Y Zou, and Yongchan Kwon. Opendataval: a unified benchmark for data valuation. *Advances in Neural Information Processing Systems*, 36, 2023.

Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Chong Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Ruslan Salakhutdinov, and Daniel Fried. Visualwebarena: Evaluating multimodal agents on realistic visual web tasks. *arXiv preprint arXiv:2401.13649*, 2024.

Ksenia Konyushkova, Raphael Sznitman, and Pascal Fua. Learning active learning from data. *Advances in neural information processing systems*, 30, 2017.

Sushant Kumar, Sumit Datta, Vishakha Singh, Sanjay Kumar Singh, and Ritesh Sharma. Opportunities and challenges in data-centric ai. *IEEE Access*, 2024.

Cecilia S Lee and Aaron Y Lee. Clinical applications of continual learning machine learning. *The Lancet Digital Health*, 2(6):e279–e281, 2020.

Hongyu Li, Liang Ding, Meng Fang, and Dacheng Tao. Revisiting catastrophic forgetting in large language model tuning. *arXiv preprint arXiv:2406.04836*, 2024.

Bing Liu, Sahisnu Mazumder, Eric Robertson, and Scott Grigsby. Ai autonomy: Self-initiated open-world continual learning and adaptation. *AI Magazine*, 44(2):185–199, 2023.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. Agentbench: Evaluating llms as agents. In *The Twelfth International Conference on Learning Representations*, 2024.

Yaxi Lu, Shenzhi Yang, Cheng Qian, Guirong Chen, Qinyu Luo, Yesai Wu, Huadong Wang, Xin Cong, Zhong Zhang, Yankai Lin, et al. Proactive agent: Shifting llm agents from reactive responses to active assistance. *arXiv preprint arXiv:2410.12361*, 2024.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*, 2023.

Curtis G Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749*, 2021.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Danrui Qi and Jiannan Wang. Cleanagent: Automating data standardization with llm-based agents. *arXiv preprint arXiv:2403.08291*, 2024.

Sumedh Rasal and EJ Hauer. Navigating complexity: Orchestrated problem solving with multi-agent llms. *arXiv preprint arXiv:2402.16713*, 2024.

Paulius Rauba, Nabeel Seedat, Max Ruiz Luyten, and Mihaela van der Schaar. Context-aware testing: A new paradigm for model testing with large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Max Read. Drowning in slop. September 2024. URL `https://nymag.com/intelligencer/article/ai-generated-content-internet-online-slop-spam.html`. Accessed: 2025-01-26.

Cedric Renggli, Luka Rimanic, Nezihe Merve Gürel, Bojan Karlaš, Wentao Wu, and Ce Zhang. A data quality-driven view of mlops. *arXiv preprint arXiv:2102.07750*, 2021.

Emma Roth. The FTC's fake review crackdown begins this fall. August 2024. URL `https://www.theverge.com/2024/8/14/24220536/ftc-fake-review-rule-ai-generated`. Accessed: 2025-01-26.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–15, 2021.

Iqbal Sarker, Alan Colman, Jun Han, and Paul Watters. *Context-aware machine learning and mobile data analytics: automated rule-based services with intelligent decision-making*. Springer Nature, 2022.

Evgeny Saveliev, Tim Schubert, Thomas Pouplin, Vasilis Kosmoliaptsis, and Mihaela van der Schaar. Climb: An ai-enabled partner for clinical predictive modeling. *arXiv preprint arXiv:2410.03736*, 2024.

Evgeny Saveliev, Jiashuo Liu, Nabeel Seedat, Anders Boyd, and Mihaela van der Schaar. Towards human-guided, data-centric llm co-pilots. *arXiv preprint arXiv:2501.10321*, 2025.

Nabeel Seedat, Fergus Imrie, and Mihaela van der Schaar. Navigating data-centric artificial intelligence with dc-check: Advances, challenges, and opportunities. *IEEE Transactions on Artificial Intelligence*, 2023.

Nabeel Seedat, Fergus Imrie, and Mihaela van der Schaar. Dissecting sample hardness: A fine-grained analysis of hardness characterization methods for data-centric ai. In *The Twelfth International Conference on Learning Representations*, 2024.

Burr Settles. Active learning literature survey. 2009.

Burr Settles. From theories to queries: Active learning in practice. In *Active learning and experimental design workshop in conjunction with AISTATS 2010*, pp. 1–18. JMLR Workshop and Conference Proceedings, 2011.

Rulin Shao, Jacqueline He, Akari Asai, Weijia Shi, Tim Dettmers, Sewon Min, Luke Zettlemoyer, and Pang Wei Koh. Scaling retrieval-based language models with a trillion-token datastore. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Kyle Swanson, Wesley Wu, Nash L Bulaong, John E Pak, and James Zou. The virtual lab: Ai agents design new sars-cov-2 nanobodies with experimental validation. *bioRxiv*, pp. 2024–11, 2024.

Wannita Takerngsaksiri, Jirat Pasuksmit, Patanamon Thongtanunam, Chakkrit Tantithamthavorn, Ruixiong Zhang, Fan Jiang, Jing Li, Evan Cook, Kun Chen, and Ming Wu. Human-in-the-loop software development agents. *arXiv preprint arXiv:2411.12924*, 2024.

Xiaqiang Tang, Jian Li, Nan Du, and Sihong Xie. Adapting to non-stationary environments: Multi-armed bandit enhanced retrieval-augmented generation on knowledge graphs. *arXiv preprint arXiv:2412.07618*, 2024.

Siying Tao, Jinyang Huang, Xiang Zhang, Xiao Sun, and Yu Gu. Dynamic memory-based continual learning with generating and screening. In *International Conference on Artificial Neural Networks*, pp. 365–376. Springer, 2023.

Megh Thakkar, Léo Boisvert, Thibault Le Sellier De Chezelles, Alexandre Piché, Maxime Gasse, Alexandre Lacoste, and Massimo Caccia. Agentmerge: Enhancing generalization in fine-tuned llm agents. In *Adaptive Foundation Models: Evolving AI for Personalized and Efficient Learning*.

Stef Van Buuren, Hendriek C Boshuizen, and Dick L Knook. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in medicine*, 18(6):681–694, 1999.

Gido M Van de Ven and Andreas S Tolias. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*, 2019.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024a.

Ruobing Wang, Daren Zha, Shi Yu, Qingfei Zhao, Yuxuan Chen, Yixuan Wang, Shuo Wang, Yukun Yan, Zhenghao Liu, Xu Han, et al. Retriever-and-memory: Towards adaptive note-enhanced retrieval-augmented generation. *arXiv preprint arXiv:2410.08821*, 2024b.

Siheng Xiong, Delin Chen, Qingyang Wu, Longxuan Yu, Qingzhen Liu, Dawei Li, Zhikai Chen, Xiaoze Liu, and Liangming Pan. Improving causal reasoning in large language models: A survey. *arXiv preprint arXiv:2410.16676*, 2024.

Frank F Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Mengxue Bao, Zora Z Wang, Xuhui Zhou, Zhitong Guo, Murong Cao, et al. Theagentcompany: benchmarking llm agents on consequential real world tasks. *arXiv preprint arXiv:2412.14161*, 2024.

Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. Data-centric artificial intelligence: A survey. *ACM Computing Surveys*, 2023.

Junhao Zheng, Chengming Shi, Xidi Cai, Qiuke Li, Duzhen Zhang, Chenxing Li, Dong Yu, and Qianli Ma. Lifelong learning of large language model based agents: A roadmap. *arXiv preprint arXiv:2501.07278*, 2025.

# Appendix - Position: What's the next frontier for Data-centric AI? Data Savvy Agents!

## Table of Contents

## A Actionable Research Directions of Data-Savvy Agent

Building on the research directions discussed in the main body, we summarize the key research pillars for each capability in Figure 2. For each capability, we categorize the detailed research directions based on their level of difficulty and novelty.

- **"Easy"** part is more closely aligned with traditional ML research and mainly involves research directions of smaller scope, such as the development of individual tools or specific functions.
- **"Medium"** part moves a step further, focusing more on the logical and reasoning aspects of data-savvy agents. This includes developing advanced strategies and building automated pipelines for tool utilization or integrating application-specific prior knowledge.
- **"Hard"** part centers on the ultimate goal of data-savvy agents—empowering them to assist and collaborate with both human experts and general AI systems. This involves enhancing their interactions with human experts and platforms, as well as enabling proactive adaptation to dynamic, ever-changing environments.

Furthermore, in addition to the research side, we would really like to highlight the other two sides, i.e. the benchmark side and the engineering side.

- **Benchmark side**: The proposed research pillars in Figure 2 are heavily dependent on specific domains and applications, making it challenging for researchers and engineers to evaluate or benchmark progress effectively. Therefore, it is crucial to develop application- and domain-specific automated benchmarks for each of these research pillars, to accelerate the development of data-savvy agents.
- **Engineering side**: The development of data-savvy agents relies heavily on the engineering side, which, much like software engineering, must ensure system stability, scalability, user experience, data privacy and security, cost efficiency, and even multi-agent collaboration. Achieving these goals requires careful design and significant effort in engineering.

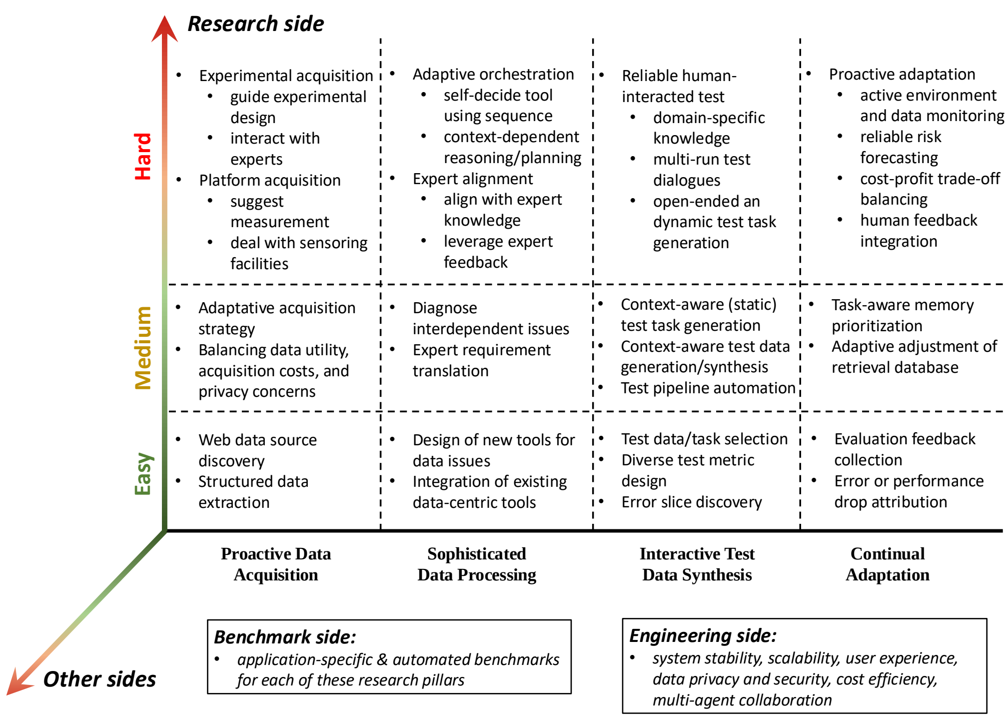


Figure 2: Summary of the actionable terms in the research of data-savvy agents.

In addition, there are several pillars not discussed in the main body, which we would like to demonstrate more here:

- **Balancing data utility, acquisition costs, and privacy concerns**: Another key challenges in real-world data acquisition is balancing the utility of the data (i.e., its ability to improve model performance) with the associated costs and privacy implications. While obtaining high-quality data from diverse sources can significantly enhance model accuracy, it often comes with high financial and logistical costs. Moreover, privacy concerns, especially with sensitive personal data, further complicate this process. For example, in healthcare or finance domains, collecting additional data to reduce model bias may conflict with data privacy regulations such as the General Data Protection Regulation (GDPR) in Europe, which aims to protect individuals' personal data and privacy, or Health Insurance Portability and Accountability Act (HIPAA) in the U.S., which sets strict standards for the handling of healthcare data. Striking a balance between maximizing data utility, minimizing costs, and ensuring privacy is a critical issue that requires careful consideration of ethical, legal, and technical factors.
- **Error slice discovery**: Error slice discovery (Eyuboglu et al., 2023; Ghosh et al., 2024; Rauba et al., 2024) involves identifying specific regions or subgroups within the data that exhibit higher error rates or performance risks. By pinpointing these areas, one can prioritize targeted data collection and implement smart deployment strategies, ultimately improving model robustness and performance in critical regions.
- **Error or performance drop attribution**: When an AI system experiences failures or performance degradation, the primary step is to identify the underlying causes. This involves attributing the performance drop to specific factors or components within the system, enabling more targeted and effective interventions to resolve the issues and enhance the system's overall performance. Recent works (Cai et al., 2023; Feng et al., 2024a) primarily focus on performance drop attribution for individual ML models. However, more principled approaches are needed to address the complexities of AI systems as a whole.

# B CURRENT DATA-CENTRIC AI VS DATA-SAVVY AGENTS

**The Shift from Static to Proactive Data-Centric AI.** The Data-Centric ML community has developed numerous approaches for improving datasets to enhance model performance. However, these methods primarily operate under a *static paradigm*, assuming a pre-defined dataset is available for processing, valuation, attribution and refinement etc.

By contrast, *Data-Savvy Agents* introduce a fundamental shift: *pro-active data-centric AI* — autonomously acquiring, refining, and adapting their data in real-time.

This represents a fundamental departure from traditional Data-Centric AI in two key ways:

- **Proactivity**: Rather than improving a fixed dataset, agents actively acquire missing data, refine existing knowledge, and generate new data points to enhance learning.
- **Adaptability**: Agents continuously adjust their data-handling strategies in response to changing environments, distribution shifts, and task requirements.

Table 1 provides a technical breakdown of key data-centric AI areas and how data-savvy agents extend them.

Table 1: Comparison of Data-Centric AI vs. Data-Savvy Agents

| Key Challenge | Research Question | Current Data-Centric AI Methods | How Data-Savvy Agents Extend This |
|---|---|---|---|
| **Data Valuation** | How do we quantify the importance of a sample for model learning? | Uses influence functions, Shapley values, and gradient-based attribution to assign importance scores. | Agents dynamically evaluate sample importance in deployment, modifying training distributions as new tasks emerge and acquiring high-value samples. |
| **Data Characterization** | Which samples are easy vs. hard to learn? | Learns sample difficulty via training dynamics, loss-based filtering, and memorization analysis. Guides data pruning and curriculum learning | Agents dynamically adjust training by acquiring new supporting data for hard examples, detecting data gaps, and modifying learning strategies in response to new challenges. |
| **Data Attribution** | How does a given sample affect model predictions? | Uses gradient-based influence estimation and feature importance methods to trace model behavior. | Agents not only trace impact but intervene, acquiring and improving data diversity dynamicallY. |
| **Active Learning** | Which unlabeled samples should be labeled next? | Uses uncertainty sampling and diversity-based selection to query labels within a predefined dataset. | Agents go beyond querying labels to identifying missing knowledge and requesting new data sources, reformulating queries dynamically. |
| **Data Cleaning & Imputation** | How do we handle noise, missingness, and inconsistencies in datasets? | Uses probabilistic imputation models and statistical heuristics, fixing errors after dataset creation. | Agents autonomously detect, verify, and correct inconsistencies by querying external sources and engaging with human experts dynamically. |
| **Distribution Shift Detection** | How do we detect changes in the data distribution over time? | Uses covariate shift detection and reweighting approaches. | Agents not only detect shifts but also autonomously adjust sampling strategies and modify data sources for retraining. |
| **Out-of-Distribution (OOD) Detection** | Can the model trust its prediction on unseen data? | Uses confidence calibration, density estimation, and contrastive learning to flag OOD samples. | Agents actively request additional evidence for uncertain inputs and retrieve external knowledge dynamically to improve robustness. |

## B.1 WHAT QUESTIONS AND RESEARCH DIRECTIONS DO DATA-SAVVY AGENTS UNLOCK?

The introduction of data-savvy agents expands the research landscape beyond what is possible with traditional Data-Centric AI approaches. We highlight the link to the main text below.

**Proactive Data Acquisition (Section 3)**:

- How can AI systems autonomously discover and integrate new knowledge?
- Traditional AI systems rely on static pre-defined datasets, whereas Data-Savvy Agents can actively seek out missing data from dynamic, multi-source environments.

- As outlined in Section 3, this enables context-aware data acquisition, where agents move beyond passive querying and actively explore, retrieve, and structure data based on evolving needs.

**Sophisticated Data Processing (Section 4)**:

- How do AI systems process and refine messy, real-world data interactively?
- Section 4 discusses the limitations of current methods, where AI agents tend to treat data processing as a fixed pipeline rather than an adaptive reasoning process.
- Data-Savvy Agents diagnose interdependent data issues, reason about missingness, and actively correct errors through self-supervised feedback loops. This unlocks new research into adaptive data reasoning, allowing AI to refine its understanding rather than merely ingesting static datasets.

**Interactive Auto-Evaluation (Section 5)**:

- Can AI systems generate their own evaluation data to test themselves?
- Current ML evaluation relies on static benchmarks, making it difficult to measure AI systems in dynamic, evolving environments.
- As outlined in Section 5, data-savvy agents introduce automated and context-aware test generation, where agents simulate interactions, generate counterfactuals, and iteratively refine their own evaluation. This enables interactive evaluation paradigms that go beyond pre-defined metrics, allowing AI to self-assess its generalization capabilities.

**Continual Adaptation (Section 6)**:

- How do AI systems remain up-to-date without full retraining?
- Static AI models fail when data distributions shift, requiring expensive re-training.
- As discussed in Section 6, current methods struggle with catastrophic forgetting and reactive rather than proactive adaptation. Data-Savvy Agents introduce incremental knowledge updating, allowing AI to retain prior knowledge while seamlessly integrating new information. This shifts towards truly lifelong learning frameworks.

Overall, Data-Savvy Agents transform AI's relationship with data, shifting from passive dataset curation to active, autonomous data reasoning. Instead of relying on static training data, these agents continuously acquire, validate, and refine their knowledge in response to real-world conditions.

### B.2 RELATED FIELDS

Beyond the literature discussed in the main body, we highlight (or call back to) several areas within the ML community that are related to *certain aspects* of the capabilities of the proposed data-savvy agent.

**Continual Learning.** Continual learning, also known as lifelong learning, refers to the ability of a model to learn continuously from new data without forgetting previously acquired knowledge. This field has seen substantial progress in recent years, with techniques designed to address challenges such as catastrophic forgetting and the integration of new knowledge over time (Van de Ven & Tolias, 2019; Lee & Lee, 2020; Wang et al., 2024a). However, while continual learning is a critical component of data-savvy agents, it cannot directly address the complex and dynamic demands of continual adaptation in data-savvy agents. The main gap lies in the difference between learning and adapting to new information. In continual learning, the focus is primarily on incremental knowledge updates within a static or predefined task space. The assumption is that the data distribution remains relatively stable and the agent's tasks are well-defined.

For data-savvy agents, however, continual adaptation involves a more proactive and flexible approach. These agents need to not only learn from new data but also autonomously adjust their decision-making processes, knowledge structures, and interactions with external systems in response to real-time changes in the environment. This includes:

1. Proactive Data Acquisition: Data-savvy agents must not only incorporate new data but also actively acquire data based on the current needs of the system, which may involve identifying gaps in knowledge or sensing when changes in the environment require adaptation.
2. Dynamic Goal Adjustment: Unlike traditional continual learning, where the learning process follows a predefined objective, continual adaptation for data-savvy agents requires frequent realignment of goals based on shifting user needs and evolving task environments.
3. Multimodal Integration: Data-savvy agents often work with diverse, real-time data sources (e.g., sensor data, user feedback, or interaction logs). Continual adaptation requires a seamless integration of this heterogeneous information into a unified model, something traditional continual learning methods often struggle with.
4. Adaptation to Changing Environments: Data-savvy agents operate in non-stationary, ever-changing environments where the data distribution, task requirements, and even the problem definitions are subject to rapid shifts. Continual learning techniques are typically not designed to handle such dynamic and unpredictable changes in real time.

Thus, while continual learning lays the foundation for maintaining and updating knowledge over time, it does not inherently account for the proactive, flexible, and context-aware adaptations required by data-savvy agents in real-world applications. Addressing this gap necessitates novel approaches that extend beyond the current scope of continual learning to incorporate real-time, context-aware adaptation and decision-making.

**Active Learning.** Please refer to the “Traditional Data Acquisition via Active Learning” in Section 3.2.

**Data-Centric Tools.** Please refer to the “Current Data-Centric Tools” in Section 4.2 as well as Table 1.

# C MORE EXAMPLES ON REAL-WORLD IMPACTS

In addition to Section 7, we put more examples on real-world impacts of data-savvy agents.

## C.1 AUTONOMOUS POLICY ADAPTATION

It is challenging for governments and global institutions to keep up with rapidly changing socio-economic and environmental situations. A data-savvy agent could transform policy-making by continuously analyzing real-time data, synthesizing insights, and generating adaptive policy recommendations. Traditional policy-making relies on slow, periodic data collection and expert analysis, making it difficult to respond quickly to crises. A data-savvy agent could proactively gather policy-relevant data, simulate different decisions, and refine recommendations based on real-world feedback.

For example, in climate policy, an AI-driven system could analyze global emissions, predict the impact of carbon reduction strategies, and adjust regulations based on new scientific findings. In economic planning, it could detect financial instability early, recommend countermeasures, and fine-tune fiscal policies in real time.

## C.2 PERSONALIZED AND LIFELONG EDUCATION

Education systems struggle to provide personalized learning at scale. A data-savvy agent could enable AI-driven lifelong education by adapting to individual learners and optimizing curricula. Traditional education relies on standardized curricula and fixed assessments, which often fail to accommodate different learning paces and styles. A data-savvy agent could continuously acquire knowledge across disciplines, update teaching strategies based on cognitive science, and personalize learning through interactive auto-evaluation. For example, in K-12 education, an AI tutor could adjust lessons in real time based on a child's progress, offering personalized exercises and explanations.

At a global level, AI-driven education could democratize access to high-quality, evolving learning resources, ensuring personalized education for all, regardless of location or background. This shift could revolutionize workforce development, accelerate innovation, and bridge global knowledge gaps.

## C.3 PRECISION HEALTHCARE

Healthcare is a data-intensive field, yet inefficiencies persist in diagnosis, treatment, and research. A data-savvy agent could revolutionize precision medicine, clinical decision-making, and drug discovery by continuously acquiring, processing, and evaluating medical data at an unprecedented scale. Traditional medical research relies on manual data collection, time-consuming clinical trials, and retrospective analysis, often leading to slow innovation cycles. A data-savvy agent could proactively acquire patient data from diverse sources (genomic data, wearable devices, EHRs, clinical studies) and generate real-time insights. Through interactive auto-evaluation, it could refine disease models, simulate drug interactions, and optimize treatment protocols based on real-world patient outcomes.

For example, in oncology, an AI-driven system could dynamically personalize cancer treatment plans by integrating real-time patient responses with the latest clinical research. In drug development, it could simulate biochemical interactions, drastically reducing the time and cost required to bring new therapies to market.

## C.4 RESILIENT GLOBAL SUPPLY CHAINS

In a volatile world, supply chains must rapidly adapt to disruptions from pandemics, geopolitics, and natural disasters. Traditional models, reliant on static data, struggle to anticipate sudden shocks. A data-savvy agent could continuously monitor trade flows, environmental conditions, and geopolitical shifts, dynamically adjusting strategies. With interactive auto-evaluation and continual adaptation, it could auto-simulate scenarios, test contingency plans, and enhance overall resilience, ensuring more responsive and robust global supply networks. For example, in global food supply chains, such a data-savvy agent could predict drought-related shortages and automatically adjust distribution to prevent famine.