

Agentic AI for Deep Data Analysis

Wachiravit Modecrua, Natchaya Temyingyong

Amity Solutions

1. Executive Summary

In today's data-intensive landscape, organizations face the challenge of extracting timely and actionable insights from increasingly complex datasets. Traditional data analysis methods are often manual, time-consuming, and struggle to scale effectively, limiting the ability to uncover the deep patterns necessary for strategic decision-making.

This white paper presents Agentic AI for Deep Data Analysis, a transformative approach that leverages autonomous AI agents to perform sophisticated, self-directed data exploration. The core of the framework lies in its ability to interpret natural language queries (NLQ) to retrieve relevant data and utilize multiple Large Language Model (LLM) agents for in-depth analysis and the generation of practical action recommendations. A critical component is Expert Alignment, which grounds the AI's processes in industry-specific knowledge and best practices, ensuring relevance, accuracy, and compliance.

The framework operates in two primary stages:

1. Data Analysis, which focuses on extracting insights from structured and external web data.
2. Action Recommendation, which generates feasible strategies based on the analysis and expert guidelines.

It addresses key challenges in data identification, understanding data structure, incorporating industry-specific compliance, and integrating external factors for a more comprehensive view.

Example applications demonstrate the framework's versatility across industries, including identifying root causes of sales decline in retail, analyzing network usage and disruptions in telecom, and understanding factors driving customer churn in business.

Initial real-world application results in a retail scenario show promising performance. Through an iterative evaluation process involving human expert baselines, "LLM as a Judge" scoring, and human re-validation, the Agentic AI's capability improved significantly, demonstrating the potential to surpass human performance levels in structured analytical tasks.

Future work will focus on:

- Building a robust **Agent Framework** with expert playbooks
- Enabling seamless analysis across mixed data sources (private, public, code-based)

- Developing a sophisticated **grading framework** with detailed metrics (Hallucination Rate, Scalability, Test Score, Overall Score),
- Validating the system against global benchmarks like the DABStep and DA-bench leaderboards.

Agentic AI for Deep Data Analysis offers a path towards more efficient, scalable, and insightful data analysis, empowering organizations to make better, data-driven decisions autonomously.

2. Introduction to Agentic AI

In the era of big data, the ability to extract meaningful insights efficiently is critical. Traditional data analysis methods often require significant human intervention, limiting scalability and adaptability. Agentic AI—an approach that leverages autonomous AI agents—offers a transformative solution by enabling self-directed, intelligent exploration of complex datasets.

By integrating agentic AI with deep data analysis, organizations can uncover hidden patterns that play a significant part for decision-making such as finding the root cause of the retail industry to improve the sales or monitoring transactions that indicate network issues in the telecom industry.

Agentic AI introduces a framework for deep data analysis across industries by integrating a Natural Language Query (NLQ) system, which converts natural language inputs into SQL queries to retrieve relevant data. Additionally, it employs multiple Large Language Model (LLM) agents to analyze data for various purposes. However, several challenges must be addressed.

Challenges:

1. **Data Identification & Retrieval:** The system must accurately determine which data should be queried and retrieved based on the input question.
2. **Understanding Data Structure:** The system must recognize the structure of the data stored in the industry's database and use the correct syntax to query it effectively.
3. **Industry-Specific Compliance:** Each industry follows specific best practices and rules that define permissible actions. The system must align with these standards to ensure compliance.
4. **Limitations of Industry Knowledge:** Relying solely on industry best practices may not cover every aspect of analysis. Incorporating external data can help expand knowledge and provide new perspectives.
5. **Incorporating External Factors:** External factors can significantly impact analysis. The system should be able to integrate relevant external data as part of the analytical process.

3. Agentic AI for Deep Data Analysis framework

Agentic AI for Deep Data Analysis is divided into two main stages: data analysis and action recommendation. The first stage focuses on analyzing the data, while the second stage provides action recommendations based on the insights derived from the first.

Each stage involves multiple agents that collaborate to analyze the data and generate the final report. This framework provides an overview, without delving into the specifics of the details of the agents working on each component. Additionally, expert alignment plays a key role across both stages, providing the foundational knowledge that guides the agents in performing tasks with the expertise necessary to ensure accuracy, consistency, and alignment with industry standards.

Expert Alignment

Expert alignment provides the foundational knowledge that guides the large language model (LLM) in performing tasks like an expert. It involves gathering insights and guidelines from industry professionals, which the LLM uses as a reference for processing data, analyzing information, and making decisions. By incorporating expert knowledge, the model understands how to approach specific tasks, ensuring its outputs align with industry standards and best practices.

This alignment helps the LLM know how to interpret complex data, solve problems, and generate actionable recommendations that reflect expert judgment and decision-making. Essentially, expert alignment serves as the base framework that directs the LLM on how to handle tasks effectively and consistently.

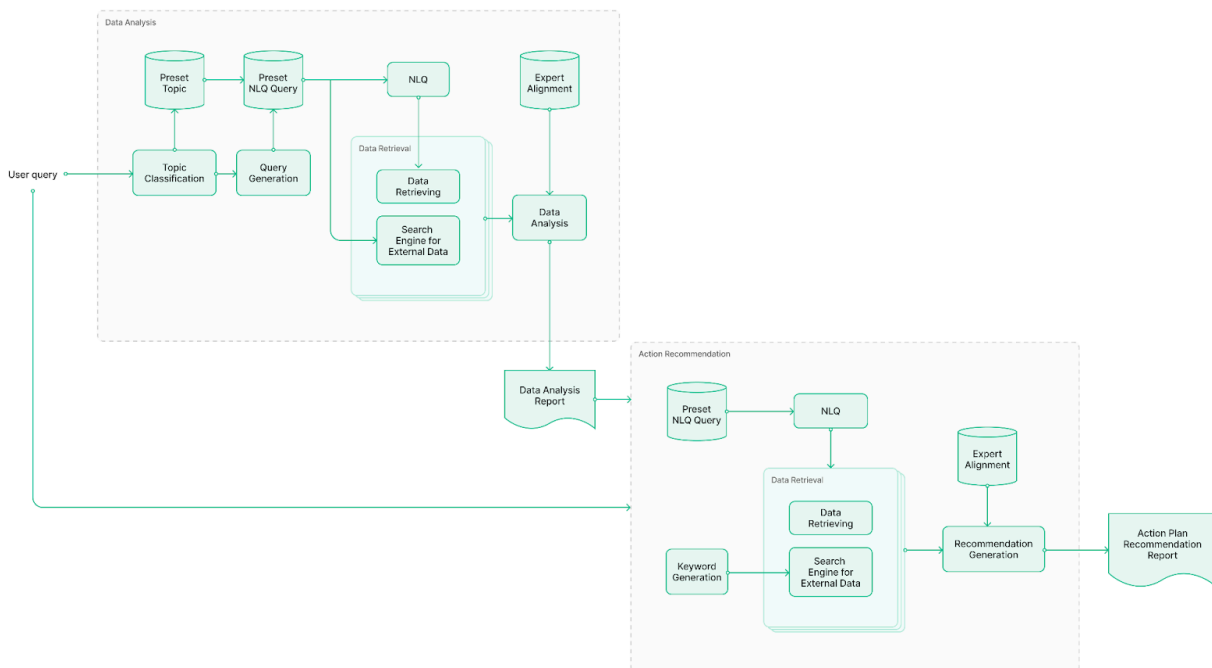


Figure 1. Agentic AI for Deep Data Analysis framework

Stage 1: Data Analysis

The data analysis process aims to extract insights from data based on the user's query. It converts natural language queries into SQL queries to retrieve structured data while also searching for relevant external factors from the internet.

To determine which data should be used, SQL query topics are preconfigured based on industry guidelines. After that, an NLQ system queries and retrieves the data, which, along with internet search results, serves as input for the data analysis component. This component leverages an LLM reasoning model to analyze the data using prompting techniques. Prompting is designed to ensure alignment with industry best practices for decision-making. The LLM then analyzes the data and generates a comprehensive report.

1. Topic classification

This is the process of identifying which topic the user's query falls under. The identified topic determines which set of preset NLQ questions should be used to retrieve the actual data, based on the guidelines associated with that topic. For example, a preset NLQ configuration includes various questions categorized by topic. Once the query is classified into Topic A, the system knows which type of NLQ questions to use based on the preset for that topic.

2. Query generation

This step involves crafting the actual queries with additional contextual information to be used in the next stages. It includes the topic identified in the previous step, along with extra context derived from a preliminary analysis of preset data. This helps refine the scope of the NLQ query and determines appropriate keywords for internet searches. For instance, in a retail root cause analysis case, this step analyzes which business channel (e.g., B2B, B2C) is experiencing the most issues. The NLQ system then scopes the query to focus on data from that specific channel.

3. Data retrieval

In this step, data is collected from two main sources: the NLQ system and web search results. The NLQ system executes structured queries to extract data from internal databases based on the refined context from the previous step. Meanwhile, a search engine powered by a large language model processes relevant keywords to gather external information from the internet. Both data sources are then combined to provide a broader context for analysis.

4. Data analysis

Once the data is retrieved, it is analyzed using a large language model (LLM) equipped with reasoning capabilities. This component takes both the structured data from the NLQ system and the external information gathered from the internet as input. Through carefully designed prompts aligned with industry best practices, the LLM evaluates the data to uncover insights, identify patterns, and support decision-making.

See more at [NLQ version 1](#)

Stage 2: Action Recommendation

Similar to the data analysis process, the grounding process involves retrieving data through the NLQ system and internet searches. Meanwhile, this stage takes the data analysis report from the first stage as input, along with the user's query, ensuring that the recommended actions align with the extracted insights.

For recommendation generation, prompting techniques are applied within the LLM reasoning model to generate suggested actions based on industry alignment guidelines. These recommendations ensure that the proposed actions are feasible and in line with industry standards.

1. Keyword generation

The keyword generation process analyzes both the user query and the data analysis report to identify key terms. These keywords are then used to find examples of successful cases or real-world implementations, providing knowledge for the LLM, which is combined with data retrieved from the NLQ system to generate action recommendations.

2. Data retrieval

Using the keywords identified in the previous step, the system searches the internet for relevant case studies, best practices, or success stories. At the same time, additional structured data may be retrieved via the NLQ system to further support the recommendation process.

3. Recommendation generation

The LLM processes all gathered information—user query, analysis report, keywords, and retrieved data—to generate actionable recommendations. These are designed to be aligned with industry standards and tailored to the specific context of the user's query and data.

4. Example Applications

Example 1: Root Cause Analysis and Recommended Actions for the Retail Industry

Analyzing the underlying causes of issues in the retail sector and recommending actions to improve sales, operations, or customer satisfaction.

User query: Why has sales in branch AA experienced a significant decline?

Root cause analysis

1. Data Analysis and Insights

- 1.1. **Sales Decline:** There is a consistent decline in sales from Q1 to Q4, with a sharp drop in Q4, the holiday season, which is typically the most profitable period for retailers.
- 1.2. **Foot Traffic Decline:** Foot traffic is also on the decline, with a 30% reduction from Q1 to Q4.
- 1.3. **Checkout Time:** A notable increase in average checkout times from 5 minutes in Q1 to 10 minutes in Q4.
- 1.4. **Customer Complaints:** The number of customer complaints grows exponentially, from 25 complaints per week in Q1 to 75 in Q4. Complaints likely pertain to long wait times, poor customer service, or inventory issues.
- 1.5. **Inventory Shortages on Popular Items:** There were stockouts on popular items from Q2 to Q4, which led to customer dissatisfaction and lost sales opportunities. By Q4, the store faces severe stockouts, especially for high-demand products.

2. Root Causes

- 2.1. **Inefficient Checkout Process:** The increase in checkout time is a significant bottleneck, particularly during peak hours in Q4. This has likely led to frustration and customers abandoning their purchases.
- 2.2. **Inventory Management Issues:** The consistent stockouts on popular items contributed to lost sales. Customers likely chose to shop elsewhere due to the unavailability of sought-after products.
- 2.3. **Declining Foot Traffic:** The 30% decrease in foot traffic over the year, especially during the crucial Q4 period, suggests that the marketing or promotional activities may not be resonating well with the target customers. Additionally, poor customer experiences (long wait times, stockouts) may have discouraged repeat visits.

Recommendations actions

1. Optimize Checkout Process

- 1.1. Increase the number of checkout staff during peak hours to reduce wait times.
- 1.2. Implement self-checkout stations to alleviate congestion at traditional checkout counters.
- 1.3. Improve point-of-sale (POS) system efficiency to speed up transactions.

2. Inventory Management

- 2.1. Implement a more robust inventory management system to predict demand and prevent stockouts, especially for high-demand products.
- 2.2. Introduce real-time stock tracking and alert systems to ensure popular items are restocked in time.

3. Customer Engagement and Marketing

- 3.1. Focus on targeted marketing campaigns, especially during peak seasons, to boost foot traffic.
- 3.2. Improve customer engagement through personalized promotions, loyalty programs, and follow-up emails to encourage repeat visits.

4. Customer Feedback System

- 4.1. Implement a better system for capturing customer feedback during and after their shopping experience. This will help to identify pain points early and enable quicker resolution.

Example 2: Data Analysis for Network Usage in the Telecom Industry

Analyzing network usage patterns to optimize performance, detect issues, and improve service quality in the telecom sector.

User query: Why is the telecom company experiencing intermittent service disruptions in certain regions?

Data Analysis

1. Network Utilization:

- 1.1. **Region A:** The network utilization is at 85%, which is relatively healthy and indicates that the network is not being overly taxed.
- 1.2. **Region B:** The network utilization is at 95%, which is quite high, suggesting that the region is under heavy load and may be experiencing congestion.

2. Service Outages:

- 2.1. **Region A:** Only 2 service outages were reported, which indicates that the network is functioning well and is able to handle the traffic without many disruptions.

- 2.2. **Region B:** A significantly higher number of service outages (15) occurred in Region B, despite the high network utilization. This suggests that the high traffic might be causing instability or that there is an infrastructure limitation contributing to these disruptions.

Possible Insights

1. **Region A:** With only 85% network utilization and minimal outages, Region A is operating under optimal conditions. The network seems to be well-equipped to handle current traffic levels, and the low number of service outages suggests **healthy network conditions**.
2. **Region B:** The high network utilization (95%) coupled with 15 service outages suggests that Region B is experiencing **network congestion**. This may be caused by either a **high user density** or **infrastructure limitations** (e.g., insufficient bandwidth, outdated hardware, etc.). The combination of high traffic and frequent outages points to a network that cannot keep up with demand, causing disruptions.

Example 3: Churn Analysis in Business

Identifying factors contributing to customer churn and recommending strategies to retain valuable customers across various business sectors.

User query: Why is the company experiencing an increase in customer churn?

Data Analysis

1. **Customer Retention Rate:**
 - 1.1. **Churn Rate:** Out of 300 customers (250 who stayed and 50 who churned), the churn rate is **16.7%** $[(50 \text{ churned} / 300 \text{ total}) * 100]$.
 - 1.2. **Subscription Revenue Loss:** The company is losing \$20 per month from each of the 50 churned customers, resulting in a **\$1,000 monthly revenue loss** from churn alone.
2. **Usage Analysis:**
 - 2.1. **Churned Customers:** Customers who churned have an average usage of **5 hours per week**, indicating low engagement with the platform.
 - 2.2. **Non-Churned Customers:** Customers who stayed on the platform have significantly higher engagement, with an average of **15 hours per week**.
 - 2.3. The substantial difference in usage between churned and non-churned customers suggests that **lower usage is a key predictor of churn**.
3. **Customer Engagement:**
 - 3.1. Customers who churned seem to be **less engaged** with the platform compared to those who stayed. This low engagement could indicate that the value or benefits of the platform are not clear to the churned customers, or they may not have had enough interaction with the platform to see its full potential.

Possible Insights

- 1. Low Engagement:** The significant drop in usage between churned customers (5 hours/week) and non-churned customers (15 hours/week) suggests that **lower usage leads to higher churn**. Customers who don't interact with the platform enough may not feel a strong connection or find enough value to continue paying for the service.
- 2. Churn Risk Factors:** The churn rate appears to be higher among customers with **less engagement**, indicating that customer engagement is a crucial factor in retention. This low engagement could be tied to the lack of onboarding, awareness of the platform's features, or underutilization of its capabilities.

5. Real-World Application Results

Root Cause Analysis and Recommended Actions for the Retail Industry

The Agentic AI framework was tested in a retail industry scenario for root cause analysis and action recommendation.

To assess the effectiveness of the Agentic AI framework in comparison to traditional human-led processes, an acceptance rate was measured at different stages:

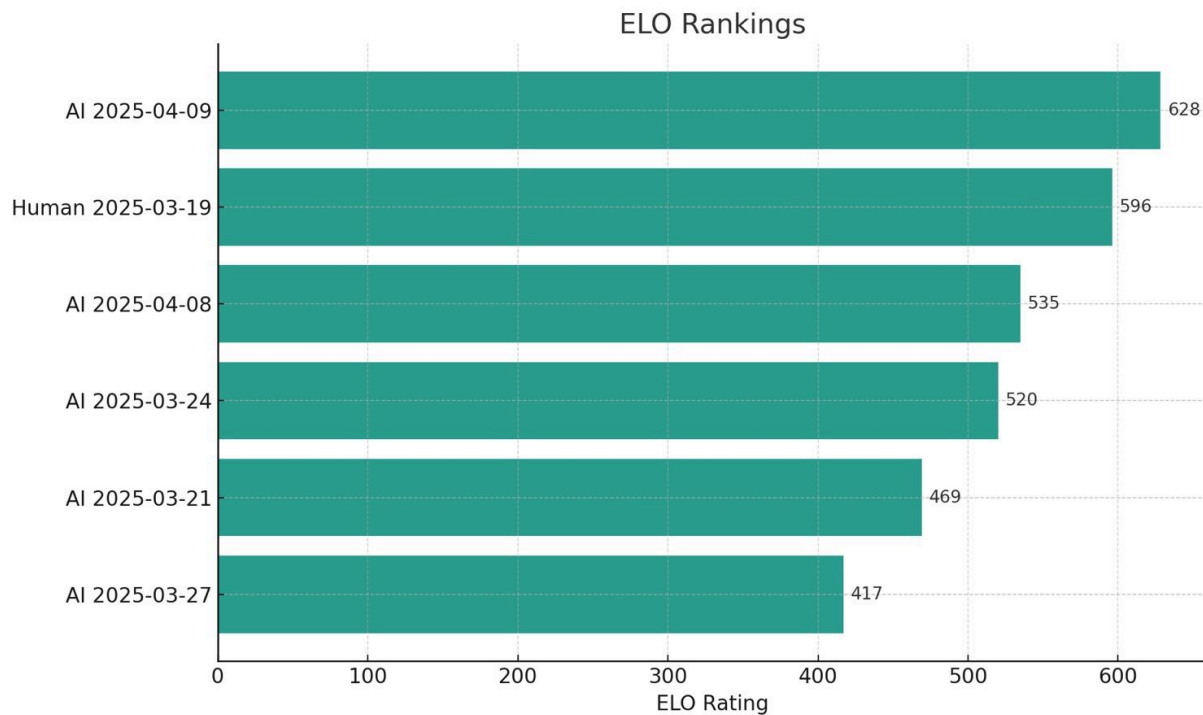


Figure 2. ELO Ranking

This evaluation framework utilizes a multi-stage process to assess and improve Agentic AI performance, integrating human expertise and automated scoring.

Use Human Expert as a Guideline

The process begins by establishing a baseline using human expert performance, as reflected in their ELO rankings (e.g., Human 2025-03-19 at 596). These rankings serve as the initial standard and target for the Agentic AI. The expertise embedded in human performance guides the development and refinement of the AI models.

Agentic AI (Initial Stage)

In its initial implementation, the Agentic AI's performance was measured, yielding ELO rankings such as AI 2025-03-27 at 417, AI 2025-03-21 at 469, and AI 2025-03-24 at 520. These initial rankings indicate that while the AI was functional, there was a performance gap compared to the human baseline, suggesting areas for improvement in its reasoning and output quality.

Scoring by “LLM as a Judge” and Re-validate by Human Expert

Following the initial assessment, an iterative process of feedback and adjustment was implemented. This involved scoring the Agentic AI's outputs, potentially using an "LLM as a Judge" mechanism for automated evaluation, followed by re-validation by human experts. This dual-validation step provided crucial insights into discrepancies and areas where the AI's performance diverged from expert expectations. The workflow was adjusted based on this feedback, including refining data inputs, improving internal processes (like keyword generation), and tuning prompting strategies to better align the AI's decision-making with expert judgment. The ELO ranking system itself reflects the outcomes of these comparisons and validations, serving as a metric for relative performance.

Agentic AI (Improved Stage)

After these significant adjustments and iterative refinement based on expert feedback and validation, the Agentic AI's performance showed marked improvement. The ELO rankings for later versions, such as AI 2025-04-08 at 535 and particularly AI 2025-04-09 at 628, demonstrate this enhancement.

The latest AI version surpasses the human baseline in ELO ranking, reflecting a significant increase in the quality and relevance of its outputs, and its ability to generate results aligned with or exceeding expert standards through this structured evaluation and improvement loop.

6. Future works

Building upon the progress in achieving high performance through iterative refinement and expert guidance, future work is centered on the development and deployment of a sophisticated Agent Framework. This framework is designed to empower the agent to perform complex data analysis tasks akin to a human expert.

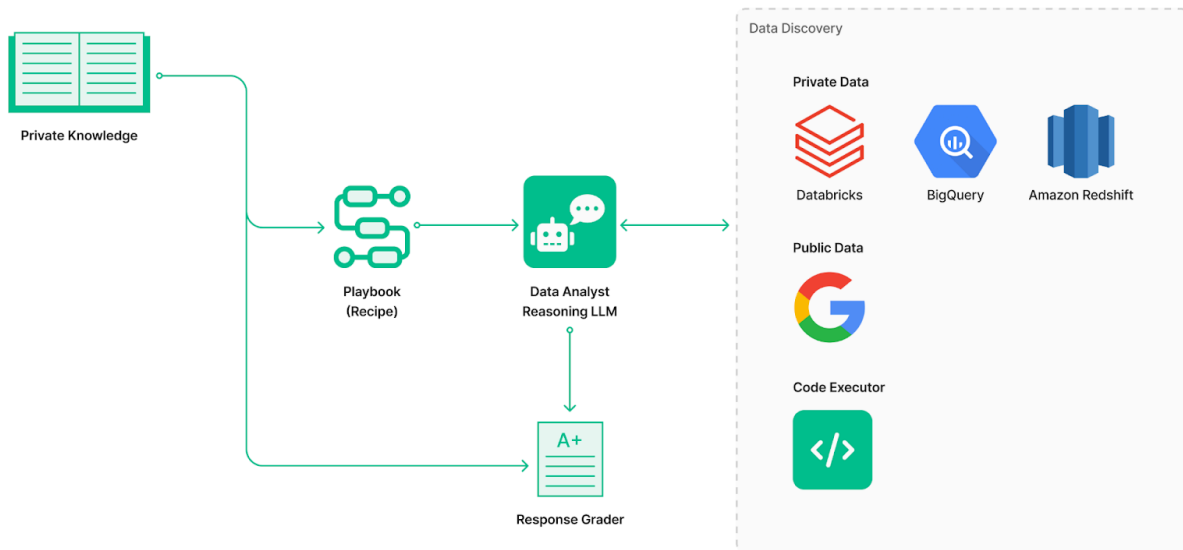


Figure 3. Sophisticated Agent Framework

1. Develop the Agent Framework and Expert Playbooks

A core focus will be the development of the Agent Framework itself, designed to operationalize expert knowledge. This involves extracting valuable insights and methodologies from human experts and translating them into actionable "Playbooks" that guide the Agent's analytical process. This directly leverages the principle of using human expertise as a fundamental guideline for AI behavior.

2. Enable Analysis Across Mixed Datasources

A key capability of the future Agent will be the ability to seamlessly analyze data from diverse sources, including Private data (potentially via a Natural Language Query - NLQ - Engine), Public data, and data requiring Code execution. This expands the Agent's versatility and applicability to real-world scenarios involving fragmented and varied data landscapes.

3. Develop a Robust Grading Framework

To ensure rigorous evaluation and continuous improvement, significant effort will be dedicated to developing a comprehensive Grading Framework. This framework will move beyond simple acceptance rates or ELO scores to incorporate detailed metrics

reflecting the quality and reliability of the Agent's output. Drawing inspiration from the provided concept, this framework will evaluate performance based on key indicators such as

- Hallucination Rate
- Scalability Score
- Test Score
- Overall Score

This framework serves as the mechanism for "Scoring by 'LLM as a Judge'" and facilitates subsequent "Re-validation by Human Expert", providing a granular assessment for targeted refinement.

4. Ensure Generalization Through Global Benchmarking

To validate the Agent's capabilities externally and ensure its performance generalizes across different tasks and datasets, the framework will be benchmarked against established Global Benchmarks. Specific targets include achieving strong performance on recognized leaderboards such as the ****DABStep Leaderboard**** and the ****DA-bench Leaderboard****. This external validation is crucial for demonstrating the Agent's effectiveness and reliability on a global scale.

By focusing on these key areas, the future work aims to create an Agentic AI capable of performing expert-level data analysis autonomously, evaluated through a rigorous grading system and validated against global standards, thereby pushing the boundaries of AI-driven insights.