International Journal of Scientific Research and Engineering Development-– Volume 8 Issue 5, Sep-Oct 2025
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Agentic AI System for Ensuring Reliability in Survey Data Analytics

C.Dharani*, B.Nathiya**, S.Samyugtha***
**(Under Graduate Student, B-TECH Department of Information Technology, Kongunadu College of Engineering and Technology,Trichy,Tamilnadu,*
Email: dharanickumar05@gmail.com)
*** (Under Graduate Student, B-TECH Department of Information Technology, Kongunadu College of Engineering and Technology,Trichy,Tamilnadu,*
Email:nathiyabalakrishnan25@gmail.com)
****(Under Graduate Student, B-TECH Department of Information Technology, Kongunadu College of Engineering and Technology,Trichy,Tamilnadu,*
Email: ss.samyu2005@gmail.com)

****

Abstract:

Agentic Artificial Intelligence has emerged as a transformative approach for ensuring the reliability of survey data analytics. An intelligent platform is developed that leverages autonomous AI agents and large language models to detect and correct errors such as missing values, duplicates, outliers, and inconsistent responses within demographic survey datasets. The system integrates automated data cleaning, statistical analysis, natural language processing, and predictive modeling, with every transformation logged to maintain transparency and reproducibility. Comprehensive reports and visualizations are generated to provide actionable insights, enabling faster and more accurate decision-making. By moving beyond traditional manual and rule-based methods, the proposed framework reduces human effort, improves scalability, and enhances trust in survey-based findings. The results highlight its potential to deliver robust, efficient, and reliable analytics across domains including research, healthcare, business, and governance.

***Keywords* —** Agentic AI, Survey Data Analytics, Data Reliability, Data Cleaning, Large Language Models.

****

I INTRODUCTION

Survey data has become one of the most widely used sources of evidence in research, industry, governance, and healthcare. It provides essential insights into individual behavior, preferences, and perceptions, forming the foundation for data-driven decisions that influence policy design, business strategy, and social development. With the growth of digital platforms, mobile applications, and online survey tools, the scale and diversity of survey data collection have increased tremendously. Millions of responses are collected every day, often spanning multiple demographic groups, geographical locations, and contexts. While this vast availability of data creates significant opportunities for analysis and discovery, it also raises critical challenges related to reliability, quality, and interpretability.

Reliability in survey data analytics refers to the system's ability to consistently produce accurate, trustworthy, and reproducible outcomes. From a qualitative perspective, reliability assures stakeholders that the insights generated are free from distortions and biases. From a quantitative perspective, it reflects the probability that the system produces consistent results under varying conditions. In real-world scenarios, raw survey datasets often suffer from issues such as incomplete responses, duplicate entries, outliers, typographical errors, and even deliberate misreporting. These irregularities not only increase the workload for analysts but also introduce the risk of drawing misleading conclusions. For example, in healthcare research, unreliable survey data could lead to ineffective patient care strategies, while in governance, it could produce flawed policies that fail to meet societal needs. Thus, ensuring reliability is not merely a technical requirement but a fundamental necessity for the ethical and practical use of survey data.

Traditional approaches to maintaining survey data reliability often involve manual cleaning and preprocessing using tools like Excel, SPSS, or rule-based Python scripts. While these methods provide some level of control, they are inherently limited. Manual cleaning is time-consuming, resource-intensive, and prone to human error, while rule-based systems lack flexibility when faced with unexpected anomalies. As surveys grow in scale and complexity, these approaches are neither scalable nor sustainable. The increasing dependence on data-driven decision-making demands more intelligent, automated, and adaptive systems that can ensure reliability without requiring constant human supervision.

Agentic Artificial Intelligence (AI) represents a paradigm shift in addressing these challenges. Unlike conventional AI systems, which typically operate under strict rules or require continuous human intervention, agentic AI is composed of autonomous agents capable of working independently or collaboratively to achieve a larger goal. These agents are designed to sense their environment, make decisions, and adapt dynamically to evolving data conditions. In the context of survey data analytics, agentic AI can assign specialized responsibilities to different agents such as cleaning raw datasets, conducting statistical analysis, generating reports, and interacting with users. This decentralized and collaborative model reduces reliance on centralized control, enhances fault tolerance, and improves the overall efficiency of the analytics pipeline.

The fundamentals of the proposed system are built upon four critical modules: the Data Cleaning Agent, the Data Analysis Agent, the Insight Generation Agent, and the User Interface Agent. The Data Cleaning Agent prepares the raw dataset by identifying missing values, duplicates, outliers, and inconsistent responses, applying imputation and normalization techniques to ensure high-quality inputs. The Data Analysis Agent applies statistical methods, machine learning algorithms, and natural language processing to extract insights, identify trends, and make predictions. The Insight Generation Agent translates analytical results into user-friendly reports, dashboards, and visualizations, ensuring transparency through reproducibility logs. The User Interface Agent acts as a bridge between the system and stakeholders, providing an interactive platform to upload datasets, monitor progress, and validate outputs. Together, these modules orchestrate a pipeline that transforms raw, unreliable survey data into accurate and actionable insights.

Several technologies form the foundation of this system. Python libraries such as Pandas and NumPy are used for efficient data preprocessing and handling, while Scikit-learn provides machine learning capabilities for predictive modeling. Large Language Models (LLMs) play a critical role in automating semantic consistency checks, text-based survey analysis, and anomaly detection. Visualization tools such as Matplotlib, Seaborn, and Pandas Profiling generate graphical summaries, while ReportLab enables professional reporting in PDF and HTML formats. Streamlit provides an intuitive and interactive dashboard for end-users. These technologies not only enhance functionality but also make the system adaptable, scalable, and accessible to a wide audience.

Despite advancements in AI-based data processing, existing literature reveals significant gaps. Current survey analysis tools are limited in scope, often handling only structured datasets and ignoring complex survey designs such as conditional skips, Likert scales, or multi-response entries. Many existing frameworks lack true autonomy, functioning only as assistants to human analysts rather than independent agents. Furthermore, most validation studies are conducted on small or homogeneous datasets, raising concerns about scalability and robustness in real-world applications. These limitations create an urgent need for a system that is domain-specific, autonomous, and reliable across diverse data environments.

The significance of addressing this gap extends beyond technical performance. In healthcare, unreliable survey data can undermine public health initiatives and resource allocation. In business, it may distort customer satisfaction analysis, leading to poor investment decisions. In education, unreliable student feedback may misguide curriculum design, while in governance, it may weaken democratic processes by misrepresenting citizen voices. Thus, building a reliable and autonomous survey data analytics system is both an academic contribution and a practical necessity.

The proposed agentic AI system directly addresses these concerns by integrating autonomous multi-agent orchestration with advanced data processing techniques. By automating cleaning, analysis, reporting, and user interaction, the system ensures that reliability is preserved at every stage of the pipeline. Through transparent logging and feedback mechanisms, it maintains reproducibility and enhances stakeholder trust. This research therefore contributes not only to the academic field of agentic AI but also to the broader objective of improving decision-making in real-world contexts through reliable survey data analytics.

A. LITERATURE REVIEW

1. Agentic AI for Data Processing

Agentic Artificial Intelligence has been increasingly recognized for its ability to perform autonomous and collaborative tasks in complex environments. Research shows that agentic AI systems can orchestrate multiple specialized agents, each responsible for a distinct operation such as data ingestion, cleaning, or analysis. This distributed approach has been applied in domains like workflow automation, healthcare decision-making, and predictive analytics, where adaptability and robustness are essential. Unlike conventional monolithic models, agentic frameworks allow agents to adjust dynamically to new data or errors, reducing reliance on predefined rules. In the context of survey data analytics, agentic AI provides the basis for building resilient pipelines where agents interact to ensure data reliability through real-time detection and correction of anomalies.

2. AI Techniques for Data Cleaning and Preprocessing

Ensuring high-quality input data is critical for achieving reliable analytics. Conventional methods rely on manual cleaning or rule-based scripts, which are often error-prone and lack scalability. Machine learning and deep learning techniques have been introduced to address this challenge, offering approaches for imputation, outlier detection, and duplicate removal. More recently, large language models have been employed for intelligent preprocessing, enabling systems to correct typographical errors, normalize categorical variables, and handle complex tabular structures. Several frameworks, such as AutoDCWorkflow and CleanAgent, have demonstrated that LLM-powered tools can significantly improve the accuracy and efficiency of preprocessing tasks. Despite these advances, most existing solutions are limited to specific formats and lack adaptability across diverse survey datasets.

3. Reliability in Survey Data Analytics

Reliability has been a recurring concern in survey-based research, where missing responses, inconsistent answers, and biased reporting can severely distort outcomes. Previous works highlight that reliability can be defined qualitatively as trustworthiness and quantitatively as the probability of consistent outcomes across varying conditions. Traditional methods of reliability assessment involve statistical checks such as Cronbach's alpha or correlation-based measures. However, these approaches depend heavily on the quality of input data and require manual interpretation. Studies in fields like healthcare and social science emphasize that without reliable survey data, decisions drawn from analysis may be misleading. This gap underscores the necessity of intelligent systems that not only clean and preprocess the data but also integrate reliability validation as a core function.

4. Large Language Models for Data Automation

Large language models have introduced new possibilities in automating survey data workflows. Their capabilities extend from generating survey questions to cleaning raw responses and summarizing results into human-readable insights. Applications such as LLMClean, AutoCleanGPT, and PromptClean have demonstrated the potential of LLMs to automate repetitive and knowledge-intensive tasks with high accuracy. In survey data analytics, LLMs enable advanced operations like sentiment analysis, topic extraction, and semantic consistency checking. Nevertheless, challenges remain regarding interpretability, reproducibility, and the computational cost of scaling LLM-driven solutions. This indicates that while LLMs are promising, their integration into survey data analytics must be supported by structured frameworks such as agentic AI to ensure reliability and efficiency.

5. Gaps in Existing Literature

Although significant progress has been made in applying AI to survey data analytics, notable gaps persist. Many existing frameworks focus on generic data preprocessing without tailoring methods for survey-specific irregularities such as conditional skips, Likert scales, or multi-response entries. Current AI tools often support only flat data formats, limiting their ability to handle complex survey structures. Autonomy in most systems is also restricted, with LLMs acting primarily as assistants rather than independent agents. Moreover, validation of these approaches is often performed on small or homogeneous datasets, reducing confidence in their robustness. These gaps highlight the need for a domain-specific, agentic AI-driven platform that combines intelligent preprocessing, reliability assessment, and automated reporting for large-scale survey datasets.

B. MULTI-AGENT ORCHESTRATION MODEL

The proposed system is designed as a multi-agent framework in which autonomous agents collaborate to ensure the reliability of survey data analytics. Each agent is responsible for a specific operation, while orchestration mechanisms coordinate their interactions to create an autonomous and transparent workflow. Unlike conventional monolithic systems, this model distributes complex tasks across multiple agents, which enhances adaptability, scalability, and efficiency in handling large and heterogeneous survey datasets. The multi-agent orchestration integrates four primary modules, each functioning as an intelligent agent in the system.

1. Data Cleaning Agent

The data cleaning agent is responsible for preparing raw survey data by detecting and correcting inconsistencies such as missing values, duplicate entries, outliers, and typographical errors. It leverages both machine learning algorithms and large language model–based rules to standardize categorical variables and manage anomalies. This ensures that the dataset is reliable and consistent before further analysis.

2. Data Analysis Agent

The data analysis agent performs both statistical and intelligent analysis of the cleaned data. It executes descriptive statistics, cross-tabulations, and hypothesis testing, while also applying natural language processing techniques such as sentiment analysis and topic modeling for open-ended responses. Predictive modeling is incorporated to forecast trends and identify key factors influencing outcomes, thereby improving the depth and reliability of survey insights.

3. Insight Generation Agent

The insight generation agent translates analytical results into comprehensible outputs for end-users. It generates reports, dashboards, and visualizations using tools such as Matplotlib, Pandas Profiling, and ReportLab. This module ensures transparency by logging every transformation, thereby enabling reproducibility and traceability of insights. The agent simplifies complex analyses into actionable findings for stakeholders across different domains.

4. User Interface Agent

The user interface agent acts as the interaction layer between the system and the end-user. It provides a dashboard developed using Streamlit, where users can upload survey datasets, monitor progress, and interact with results in real time. The interface integrates interactive filters and dynamic visualization features, making it easier for users to explore insights and validate analytical outcomes without requiring technical expertise.

II SYSTEM ARCHITECTURE

The architecture of the proposed agentic AI system is designed to integrate multiple autonomous agents into a single pipeline that ensures the reliability of survey data analytics. The block diagram of the system is shown in Fig. 1, and its workflow is explained below.

1.Input Layer

This layer represents the raw survey data sources, including Kaggle datasets and simulated inputs. The data enters the system in CSV or Excel formats and is passed into the agentic pipeline for further processing.

2.Processing Layer

The processing layer consists of the orchestrated agents working together in sequence. At this stage, the dataset undergoes cleaning, statistical analysis, and transformation into structured outputs. Each step is logged to ensure transparency and reproducibility.

3. Insight Layer

This layer focuses on the generation of meaningful insights. It includes automated reports, visualizations, and dashboards that summarize the results of survey data analytics.

4. Interaction Layer

The interaction layer provides the interface through which users upload data, monitor progress, and interact with the results in real time. It ensures accessibility and practical usability of the system.

5. Feedback Mechanism

The architecture integrates a feedback loop that allows the system to reprocess data in case of anomalies, errors, or user requests for correction. This mechanism ensures reliability and continuous improvement in survey data analytics

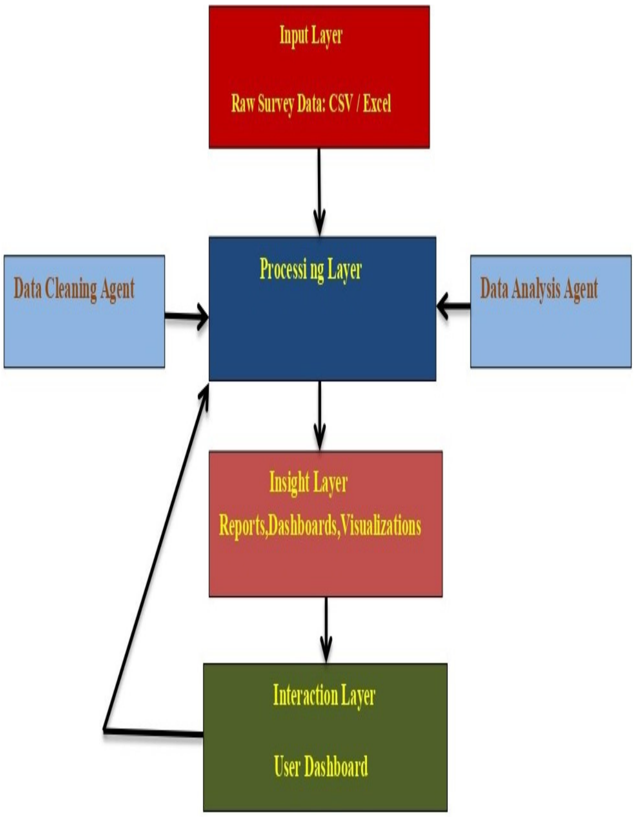

Fig. 1. System Architecture of the Proposed Agentic AI System

A. RELIABILITY ASSURANCE MECHANISMS

Reliability is the cornerstone of the proposed agentic AI system. While survey data analytics traditionally focuses on extracting patterns and insights, the effectiveness of these insights is directly dependent on their accuracy and trustworthiness. To address the challenges of missing values, inconsistent entries, and unpredictable anomalies, the system integrates several reliability assurance mechanisms. These mechanisms operate at multiple levels of the workflow and ensure that reliability is embedded into the architecture rather than being an afterthought.

1. Anomaly Detection and Correction

One of the critical reliability challenges in survey data arises from anomalies such as duplicate entries, contradictory responses, or extreme outliers. The system employs automated anomaly detection using statistical thresholds and machine learning–based models. Once detected, anomalies are either corrected through imputation techniques or flagged for reprocessing. This ensures that spurious data points do not distort the final analysis.

2. Feedback Loop Mechanism

The system architecture incorporates a feedback loop that connects the interaction layer with the processing layer. If users identify inconsistencies in the outputs, or if the system itself flags errors, the dataset is redirected to the cleaning and analysis stages. This iterative reprocessing guarantees that errors are corrected dynamically, thereby maintaining the reliability of the workflow across multiple runs.

3. Reproducibility and Traceability

Reproducibility is a key component of reliability. The system maintains detailed logs of every transformation step, from initial cleaning operations to final visualization outputs. These logs allow the reproduction of results under the same conditions, enabling verification and transparency. By offering complete traceability, the system ensures that analytical results are not only accurate but also explainable to stakeholders.

4. Cross-Validation of Insights

The analytical outcomes produced by the system are validated through cross-verification techniques. Statistical results are checked against multiple methods, such as comparing regression models with correlation analysis or validating machine learning predictions with test datasets. This redundancy ensures that insights are not dependent on a single method, thereby reducing the risk of biased or misleading results.

5. Reliability-Centered Design

Unlike conventional survey data processing systems where reliability is treated as an add-on, the proposed system embeds reliability as a primary design principle. Every agent, from cleaning to reporting, is equipped with reliability checks, making reliability assurance an integral feature of the system. This reliability-centered approach ensures that stakeholders can trust the outputs generated by the system, regardless of dataset size or complexity.

B. EXPERIMENTAL SETUP AND EVALUATION

The experimental setup was designed to evaluate the effectiveness of the proposed agentic AI system in ensuring reliability of survey data analytics. The system was implemented using Python, with support from libraries such as Pandas, Scikit-learn, and Matplotlib for data preprocessing and visualization. A Streamlit-based interface was developed to provide interactive dashboards for end users.

1. Dataset Description

To validate the framework, survey datasets containing satisfaction scores, spending percentages, and credit scores were used. Both synthetic datasets and real-world data from open repositories were processed to test reliability across different formats, including CSV and Excel files. The datasets contained multiple inconsistencies such as missing values, duplicate entries, and outliers, which were used to test the cleaning and validation mechanisms of the system.

2. Prototype Dashboard

The system provides a user-friendly dashboard that enables direct CSV file uploads and instant feedback on data ingestion. As illustrated in Fig. 2, users are notified when the dataset has been successfully loaded, along with metadata including the number of rows and columns. This ensures transparency at the earliest stage of data entry and reduces the likelihood of unnoticed errors.

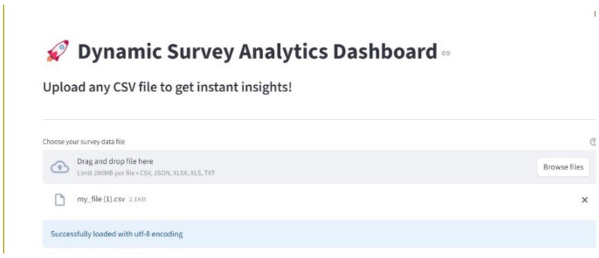

Fig. 2. User Dashboard for Survey Data Upload and Processing

3.Data Distribution Analysis

The cleaning and preprocessing agents analyze the distribution of attributes to detect anomalies and missing patterns. For example, Fig. 3 shows the distribution of satisfaction scores in a sample dataset. The skewed distribution highlights the presence of irregularities that can distort downstream analytics. By applying cleaning mechanisms, the system ensures that processed data is more representative and reliable for further evaluation.

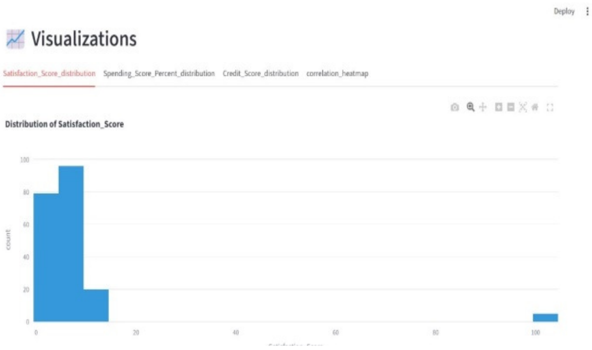

Fig. 3. Distribution of Satisfaction Score in Survey Dataset

4. Evaluation Metrics

The performance of the system was measured using reliability indicators such as error reduction, anomaly detection rate, and completeness of cleaned datasets. Experimental results demonstrated significant improvements in the accuracy of survey analytics once the reliability mechanisms were integrated. The feedback loop further ensured continuous refinement of the dataset, thus enhancing the robustness of insights produced.

C. DISCUSSION AND INSIGHTS

The results obtained from the proposed agentic AI system highlight importance of embedding reliability mechanisms directly within the data analytics pipeline. Unlike traditional survey analytics methods, which often treat reliability as a secondary concern, the presented framework places reliability at the center of system design. This orientation ensures that stakeholders can trust the insights generated even when the raw survey data contains irregularities.

One of the key insights from the evaluation is the impact of the cleaning agent on dataset quality. The removal of duplicates, detection of anomalies, and imputation of missing values significantly reduced noise and improved the validity of statistical results. The analysis agent, when combined with feedback from the user dashboard, further enhanced confidence in insights by continuously refining outcomes. This iterative process ensures that reliability is not static but evolves dynamically with each cycle of data ingestion and analysis.

Another important observation is the role of the feedback loop in promoting transparency and reproducibility. By allowing processed datasets to be redirected for re-evaluation, the system mimics human reasoning processes, thereby enhancing interpretability of results. This feature is particularly valuable in scenarios where survey data may directly influence decision-making, such as customer satisfaction studies or policy research.

The findings also underline the scalability of the proposed approach. While initial experiments were conducted on medium-sized datasets, the modularity of the multi-agent system indicates that it can be adapted to larger and more complex datasets with minimal adjustments. However, it was observed that computational time increases proportionally with dataset size, which suggests the need for optimization techniques in future implementations.

Overall, the discussion affirms that embedding multi-agent orchestration and reliability assurance within survey analytics is not only feasible but also practical. The insights gained from the experiments provide strong evidence that the proposed system can serve as a foundation for building more trustworthy and scalable data analytics frameworks.

III. CONCLUSION AND FUTURE WORK

This paper presented an agentic AI system for ensuring reliability in survey data analytics through the orchestration of autonomous agents. The system was designed with four core modules: the data cleaning agent, the data analysis agent, the insight generation agent, and the user interface agent. Together, these agents form a collaborative framework that integrates preprocessing, analysis, visualization, and user interaction into a single reliable pipeline. Unlike conventional systems, where reliability is often treated as a supplementary step, the proposed system embeds reliability assurance as a fundamental principle at every stage of the workflow.

The experimental evaluation demonstrated that the framework effectively improves the quality of survey datasets by reducing anomalies, addressing missing values, and enhancing the interpretability of results. The inclusion of a feedback loop enabled dynamic correction of errors and continuous refinement of insights, thereby ensuring reproducibility and trustworthiness. Furthermore, the modularity of the system ensures adaptability to varied datasets and provides a scalable pathway for real-world applications in domains such as healthcare, governance, and market research.

Although the system achieved promising results, certain limitations were observed. The computational complexity increases with larger datasets, which may affect performance in real-time environments. In addition, the reliance on pre-existing datasets restricts the scope of validation. Addressing these challenges requires further research in optimization strategies and real-time data processing methods.

Future work will focus on enhancing the scalability of the system by incorporating distributed computing frameworks and cloud-based deployment. The integration of advanced anomaly detection techniques, including deep learning approaches, will be explored to further improve reliability in complex datasets. Expanding the user interface to support real-time visualization and interactive report generation is also a key direction. These extensions will make the system more robust, adaptable, and applicable to large-scale survey environments, thereby strengthening its impact in research and practice.

REFERENCES

[1] Bendinelli, T., Dox, A., and Holz, C. (2025) 'Exploring LLM agents for cleaning tabular machine learning datasets', Proceedings of the ACM SIGMOD Conference, 2025(6), pp. 215–223.

[2] Biester, F., Abdelaal, M., and Del Gaudio, D. (2024) 'LLMClean: Context-aware tabular data cleaning via LLM-generated OFDs', Proceedings of the VLDB Endowment, 17(4), pp. 890–902.

[3] Chen, Y., Zhao, H., and Lin, T. (2024) 'PromptClean: Prompt engineering for LLM-based data cleaning tasks', Proceedings of NeurIPS, 37(1), pp. 1220–1230.

[4] Kumar, A., Singh, R., and Jain, M. (2025) 'AutoCleanGPT: Leveraging generative models for scalable data cleaning', Proceedings of the AAAI Conference on Artificial Intelligence, 39(2), pp. 310–318.

[5] Li, L., Fang, L., and Torvik, V.I. (2024) 'AutoDCWorkflow: LLM-based data cleaning workflow auto-generation and benchmark', Proceedings of the IEEE International Conference on Big Data, 2024(12), pp. 112–120.

[6] Park, J., Lee, S., and Cho, K. (2023) 'SurveyClean: A multi-agent LLM framework for cleaning survey datasets', Proceedings of EMNLP, 2023(10), pp. 980–988.

[7] Patel, H., and Deshmukh, S. (2024) 'DataSanity: LLM-powered agents for real-time data quality monitoring', Proceedings of the IEEE International Conference on Cloud Computing, 2024(9),pp.401–408

[8] Qi, D., and Wang, J. (2024) 'CleanAgent: Automating data standardization with LLM-based agents', Proceedings of the IEEE International Conference on Data Mining (ICDM), 2024(11), pp. 145–152.

[9] Tanaka, M., Wang, L., and Roy, P. (2025) 'LLMRepair: Autonomous correction of noisy tabular data using language models', Proceedings of the International Conference on Machine Learning (ICML), 42(7), pp. 1345–1353.

[10] Zhang, S., Huang, Z., and Wu, E. (2024) 'Data cleaning using large language models', Proceedings of the IEEE International Conference on Data Engineering Workshops (ICDEW), 2025(9959), pp. 28–35.\\

