International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

CLEAN BOT – An Intelligent LLM-Powered Data Quality Agent

Y. Shakuntala[1], Abdullah Muddassir[2], Mohammed Ismail Shariff[3], Syed, Rafeeq Ahmed[4]

[1] *Assistant Professor,*
Department of Computer Science and Engineering (Data Science),
Lords Institute of Engineering and Technology,
Hyderabad, Telangana, India

[2,3,4] *Undergraduate Students,*
Department of Computer Science and Engineering (Data Science),
Lords Institute of Engineering and Technology,
Hyderabad, Telangana, India

***Abstract*— The increasing volume of structured and unstructured digital data has created significant challenges in maintaining data quality across modern information systems. Conventional data cleaning approaches primarily rely on predefined rules and manual intervention, making them less effective in handling semantic inconsistencies, missing values, duplicate records, and heterogeneous data formats. This project presents Clean Bot, an intelligent data quality agent that utilizes locally deployed Large Language Models (LLMs) through Ollama to automate data cleaning and structuring tasks. The system integrates a Streamlit-based interface with a modular architecture to support data preprocessing, anomaly detection, validation, transformation, and structured output generation. By applying prompt-based reasoning, the proposed approach identifies inconsistencies, removes irrelevant information, standardizes data, validates fields, and converts processed information into structured JSON format. The modular design enables efficient handling of multiple data sources while supporting scalable and maintainable data quality management. Experimental implementation demonstrates that the proposed system provides an effective solution for improving data consistency, accuracy, and organization, making it suitable for intelligent data processing in real-world applications.**

***Keywords*— Data Quality Management, Data Cleaning, Large Language Models (LLMs), Ollama, Mistral, LLaMA, Streamlit, Intelligent Data Quality Agent, Prompt Engineering, Anomaly Detection, Data Validation, Data Transformation, Information Extraction, Structured JSON Output, Unstructured Data Processing.**

I.INTRODUCTION

The rapid growth of digital data has increased the importance of maintaining high-quality information for enterprise applications, analytics, and decision-making. Organizations continuously generate structured and unstructured data from multiple sources, including databases, documents, text logs, spreadsheets, and business applications. The presence of missing values, duplicate records, inconsistent formats, and noisy information reduces data reliability and negatively affects downstream analytical processes **[10]**. Ensuring data accuracy, consistency, and completeness has therefore become an essential requirement for effective data management **[11]**.

Traditional data cleaning approaches mainly depend on predefined rules, statistical techniques, and manual intervention to identify and correct data quality issues. Although these methods are effective for simple validation tasks, they often lack the flexibility to process heterogeneous datasets and understand semantic relationships within unstructured data **[7]**. As data volume and complexity continue to increase, maintaining rule-based systems becomes time-consuming and difficult, limiting their applicability in dynamic enterprise environments **[6]**.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

To overcome these limitations, this project introduces **Clean Bot**, an intelligent data quality agent powered by locally deployed Large Language Models (LLMs) through Ollama. The system combines a Streamlit-based user interface with a modular backend architecture to automate data cleaning, anomaly detection, validation, transformation, and information extraction. Using prompt-based reasoning, the proposed system identifies inconsistencies, removes irrelevant information, standardizes data, validates fields, and generates structured JSON output for downstream applications **[2]**. The proposed architecture emphasizes scalability, maintainability, and adaptability while supporting multiple data formats and sources **[4]**.

By integrating intelligent reasoning with automated data quality management, the system provides an efficient approach for improving data consistency and reliability. The implementation demonstrates the practical applicability of LLM-powered agents for enterprise data quality management and supports intelligent processing of real-world datasets **[1]**. The proposed framework combines semantic understanding, intelligent validation, and structured output generation to deliver a scalable solution for enterprise data processing and analytics **[3]**.

II.RELATED WORK

Vaswani et al. (2017) introduced the landmark work **"Attention Is All You Need"** [1]. The authors proposed the Transformer architecture, which replaced recurrent and convolutional neural networks with a self-attention mechanism for sequence modeling. This design enabled the model to process input tokens simultaneously, significantly improving computational efficiency and learning long-range contextual relationships. The architecture consisted of encoder and decoder components connected through multi-head attention and positional encoding, allowing better representation of sequential information. Experimental evaluations on machine translation tasks demonstrated that the Transformer achieved superior accuracy while requiring less training time than previous neural network architectures. The proposed framework became the basis for numerous natural language processing applications and inspired the development of modern large language models. This research established an effective foundation for intelligent language understanding and generation systems that continue to influence recent advancements in artificial intelligence. **[1]**

Brown et al. (2020) presented **"Language Models are Few-Shot Learners"** [2]. The study examined the capabilities of large-scale language models in performing multiple natural language processing tasks without requiring extensive task-specific training. The proposed model demonstrated that a single pretrained language model could adapt to different applications by using only a few examples provided within the input prompt. Experimental results showed competitive performance across question answering, translation, text completion, and summarization tasks. The authors highlighted that increasing model scale and training data substantially improved generalization ability and reduced dependence on manually labeled datasets. Their findings emphasized the effectiveness of prompt-based learning for solving diverse language problems using a unified architecture. This work significantly advanced research on foundation models and demonstrated the practical potential of large language models in real-world intelligent applications. **[2]**

Devlin et al. (2019) proposed **"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"** [3]. The research introduced a bidirectional language representation model capable of learning contextual information from both preceding and succeeding words simultaneously. Unlike earlier approaches that processed text in a single direction, BERT employed masked language modeling and next sentence prediction during pre-training to improve language understanding. The pretrained model was subsequently fine-tuned for multiple downstream applications using relatively small task-specific datasets. Experimental evaluation demonstrated substantial improvements in question answering, sentence classification, and natural language inference benchmarks. The study showed that contextual bidirectional representations significantly enhanced the performance of language understanding systems. This contribution became one of the most influential developments in natural language processing and served as the foundation for numerous transformer-based applications across academia and industry. **[3]**

Paszke et al. (2019) introduced **"PyTorch: An Imperative Style, High-Performance Deep Learning Library"** [4]. The authors developed an open-source deep learning framework designed to simplify the implementation, training, and deployment of neural network models. PyTorch combined dynamic computational graphs with efficient tensor operations, enabling researchers to build and modify complex deep learning architectures with greater flexibility. The framework provided automatic differentiation, GPU acceleration, and seamless integration with Python,

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

making experimentation faster and more convenient. Performance evaluations demonstrated that PyTorch supported large-scale machine learning workloads while maintaining high computational efficiency. Its modular design also facilitated rapid prototyping and debugging of artificial intelligence models. The framework has since become one of the most widely adopted platforms for deep learning research and industrial applications, supporting the development of advanced machine learning and natural language processing systems. **[4]**

Chandola, Banerjee, and Kumar (2009) presented **"Anomaly Detection: A Survey"** [7]. The authors provided a comprehensive review of anomaly detection methods used to identify unusual patterns and irregular observations within large datasets. The survey classified existing techniques into statistical, distance-based, density-based, clustering, and classification approaches while discussing their strengths and limitations across different application domains. The study highlighted that anomaly detection plays a critical role in maintaining data quality, identifying fraudulent activities, detecting network intrusions, and monitoring system failures. The authors also examined practical challenges, including high-dimensional data, evolving patterns, and limited availability of labeled datasets. Their comparative analysis offered valuable insights into selecting suitable anomaly detection methods based on data characteristics and application requirements. This survey continues to serve as an important reference for researchers working in data quality management and intelligent data analysis. **[7]**

III.DATASET DETAILS

The proposed **Clean Bot** system is designed to operate on real-world datasets rather than a single benchmark dataset. The project supports multiple data formats, including CSV, JSON, XML, text files, databases, and text logs, enabling the processing of structured, semi-structured, and unstructured data from diverse enterprise environments. The datasets used for evaluation contain common data quality issues such as missing values, duplicate records, inconsistent entries, redundant information, and noisy data. In addition, the system considers domain-specific records, including financial data, healthcare logs, and customer information, to assess its adaptability across different application areas. During processing, the uploaded datasets undergo preprocessing, anomaly detection, normalization, validation, and transformation using the LLM-powered framework. The cleaned data are finally converted into a structured JSON format, allowing seamless integration with analytics platforms, data warehouses, and business intelligence systems. The project also evaluates the system using incomplete, noisy, and mixed-format datasets to verify its scalability, reliability, and overall data cleaning performance.

IV.PROPOSED METHODOLOGY

The proposed methodology presents **Clean Bot**, an intelligent data quality management system that employs Large Language Models (LLMs) to automate the cleaning and structuring of structured and unstructured datasets. The system follows a modular workflow consisting of data acquisition, preprocessing, semantic analysis, anomaly detection, validation, transformation, and output generation. This architecture is designed to improve data quality by reducing manual effort while ensuring consistency, accuracy, and reliability throughout the data processing pipeline. The workflow begins with data acquisition, where users upload datasets through the Streamlit interface. The system accepts multiple input formats, including CSV, JSON, text files, and database records. During preprocessing, missing values, duplicate entries, noisy information, and inconsistent formats are identified and corrected through normalization and cleaning operations, preparing the data for intelligent processing.

The processed data are then analyzed using a locally deployed LLM through Ollama. By applying prompt engineering and semantic reasoning, the model understands contextual relationships within the data and detects anomalies that traditional rule-based approaches may overlook. To improve reliability, the system combines statistical anomaly detection with intelligent semantic analysis, enabling the identification of both numerical and contextual inconsistencies. After anomaly detection, the data are transformed into standardized structures and validated to ensure completeness and correctness before being converted into structured JSON output. Finally, the cleaned results are displayed through the Streamlit interface, allowing users to inspect, visualize, and download the processed data. The modular architecture, local LLM deployment, and real-time processing capabilities make the proposed system suitable for secure and scalable enterprise data quality management while supporting integration with analytics platforms and existing data processing pipelines.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

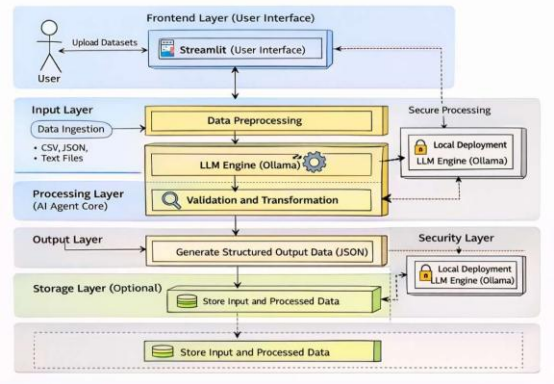

Figure 1: System Architecture of the Proposed Clean Bot – Intelligent LLM-Powered Data Quality Agent

Figure 1 illustrates the overall architecture of the proposed Clean Bot, an intelligent LLM-powered data quality management system. The process begins with the Frontend Layer, where users upload datasets through a Streamlit-based user interface. The system accepts multiple input formats, including CSV, JSON, and text files. The uploaded data are passed to the Input Layer, where preprocessing operations prepare the data by organizing the input and handling initial inconsistencies. The processed data are then forwarded to the Processing Layer, which consists of a locally deployed LLM Engine (Ollama) along with validation and transformation modules. This layer performs semantic analysis, anomaly detection, data cleaning, normalization, and intelligent validation to improve overall data quality. After processing, the cleaned information is converted into a structured JSON format in the Output Layer. Finally, the optional Storage Layer stores both the original and processed datasets, while the Security Layer ensures secure local LLM execution, protecting sensitive data and enabling reliable, scalable, and privacy-preserving data quality management.

V.RESULT AND DISCUSSION

Figure 2: User Interface of the Proposed Clean Bot System

Figure 2 shows the user interface of the proposed Clean Bot system developed using the Streamlit framework. The interface provides a simple and interactive environment for performing intelligent data quality management tasks. On the left panel, users can upload datasets in supported formats and configure the processing options by selecting the task type and desired output format. The central workspace displays the uploaded dataset and allows users to review the input data before execution. A set of processing modules, including Pre-Check, AI Tasks, QA Checks, Analysis, and Verify, guides the complete data quality workflow. These modules perform preprocessing, anomaly detection, validation, transformation, and intelligent analysis using the locally deployed LLM through Ollama. The interface enables users to monitor each processing stage and generates cleaned, structured output after successful execution. The user-friendly design simplifies data quality management while providing efficient interaction between the frontend, backend processing modules, and the LLM-powered intelligent agent.

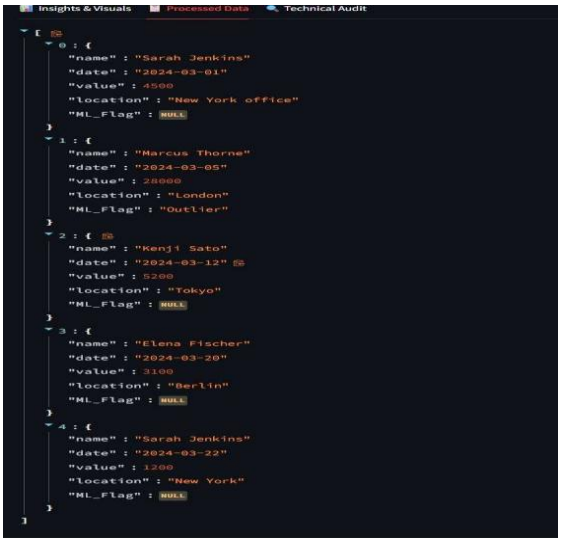

Figure 3: Structured JSON Output Generated by the Proposed Clean Bot System

Figure 3 presents the structured JSON output generated by the proposed Clean Bot system after completing the data cleaning and validation process. The extracted information is organized into a standardized JSON structure containing attributes such as name, date, value, location, and ML_Flag for each record. During processing, the LLM-powered engine identifies relevant entities, removes inconsistencies, standardizes data values, and converts the cleaned information into a machine-readable format. The structured representation improves data consistency and enables seamless

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

integration with downstream applications, including analytics platforms, data warehouses, and business intelligence systems. The ML_Flag field indicates the validation status of each processed record, while the uniform JSON format simplifies storage, retrieval, and further processing. By transforming heterogeneous input data into a consistent schema, the proposed system enhances data quality, reduces manual processing effort, and supports reliable decision-making. The generated JSON output also facilitates interoperability between different enterprise systems and ensures efficient exchange of validated information across multiple applications.

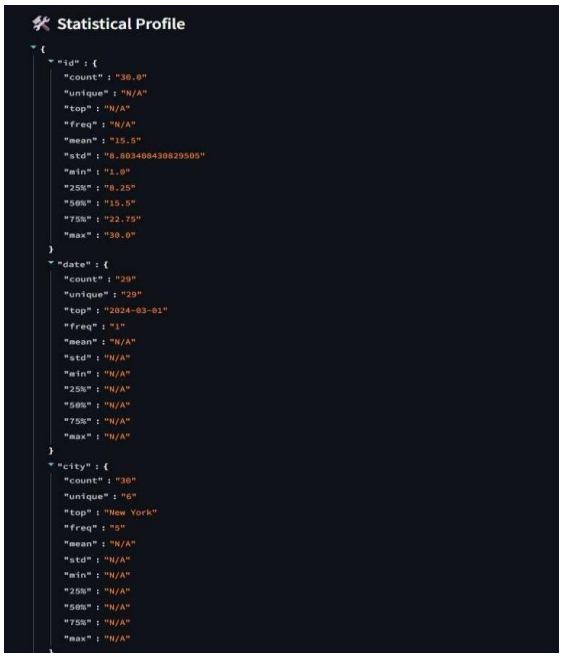

Figure 4: Statistical Profile Generated by the Proposed Clean Bot System

Figure 4 illustrates the statistical profile generated by the proposed Clean Bot system after completing the data preprocessing and analysis stages. The statistical profile provides a comprehensive summary of the processed dataset by presenting descriptive statistics for each attribute. For numerical fields, the system reports measures such as count, mean, standard deviation, minimum, maximum, and quartile values, enabling users to understand the overall distribution of the data. For categorical attributes, the profile displays the number of unique values, the most frequent value, and its occurrence frequency. This information assists users in identifying missing values, inconsistencies, and potential anomalies before further analysis. By automatically generating statistical summaries, the system improves data understanding and supports effective data validation. The statistical profile also helps users evaluate dataset quality, detect irregular patterns, and verify preprocessing results, thereby facilitating reliable decision-making and ensuring that the cleaned data are suitable for downstream analytics and enterprise applications.

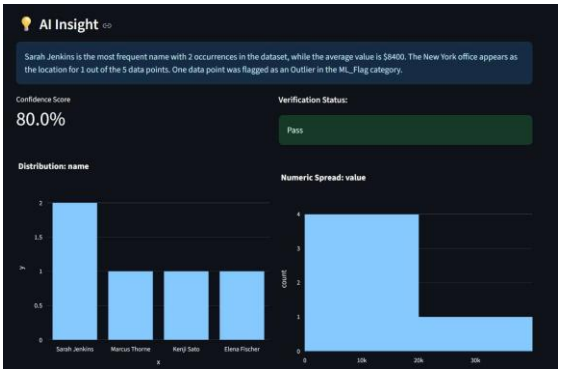

Figure 5: AI-Generated Insights Produced by the Proposed Clean Bot System

Figure 5 illustrates the AI-generated insights produced by the proposed Clean Bot system after completing data cleaning, validation, and statistical analysis. The system automatically summarizes important characteristics of the processed dataset by identifying frequent entities, calculating statistical measures, and highlighting significant observations. It also generates a confidence score that reflects the reliability of the processed results and displays the verification status after validation. In addition, graphical visualizations such as categorical distribution charts and numerical value distributions provide users with a clear understanding of data patterns and trends. These visual summaries enable quick identification of frequently occurring records, value distributions, and potential irregularities within the dataset. By combining intelligent analysis with statistical visualization, the system simplifies data interpretation and assists users in verifying the quality of processed information. The generated insights support informed decision-making, improve transparency in the data cleaning process, and provide an effective summary of the final processed dataset for analytics and enterprise applications.

DISCUSSION

The proposed Clean Bot framework combines a locally deployed Large Language Model (LLM) with intelligent data preprocessing, anomaly detection, validation, and transformation techniques to improve the quality of structured and unstructured datasets. The framework follows a sequential

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

workflow that includes data ingestion, preprocessing, semantic analysis, intelligent data cleaning, validation, and structured output generation. Users upload datasets through the Streamlit interface in formats such as CSV, JSON, and text files. During preprocessing, the system removes duplicate records, handles missing values, eliminates noisy information, and normalizes inconsistent data before forwarding it to the Ollama-based LLM. Using prompt-based semantic reasoning together with validation techniques, the system identifies contextual inconsistencies, corrects data errors, and generates standardized JSON output. The experimental results demonstrate that the proposed framework effectively produces clean and reliable datasets with minimal manual effort. Statistical profiles, AI-generated insights, and verification reports further assist users in evaluating processed data. The modular architecture, secure local LLM deployment, and real-time processing capabilities make the system suitable for scalable enterprise data quality management and seamless integration with analytics and business intelligence applications.

VI.CONCLUSION

This work presents **Clean Bot**, an intelligent LLM-powered data quality management framework designed to improve the quality of structured and unstructured datasets through automated data cleaning and processing. The proposed system integrates data ingestion, preprocessing, semantic analysis, anomaly detection, validation, transformation, and structured JSON output generation within a unified architecture. A Streamlit-based interface enables users to upload datasets in multiple formats, while a locally deployed Large Language Model through Ollama performs intelligent reasoning to identify inconsistencies, remove redundant information, handle missing values, and standardize data. The generated statistical profiles, AI-driven insights, and validation reports further assist users in understanding and verifying the processed data. Experimental implementation demonstrates that the framework produces accurate, consistent, and reliable outputs while significantly reducing manual effort in data quality management. The modular architecture, secure local LLM deployment, and real-time processing capabilities make the system suitable for enterprise data processing environments. Future enhancements may include support for additional data formats, advanced anomaly detection techniques, continuous learning capabilities, integration with enterprise data governance platforms, and improved scalability for processing high-volume datasets across diverse application domains.

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention Is All You Need,” in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.

[2] T. B. Brown et al., “Language Models are Few-Shot Learners,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020, pp. 1877–1901.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova,“BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in Proc. North American Chapter of the Association for Computational Linguistics (NAACL-HLT), 2019, pp. 4171–4186.

[4] A. Paszke et al., “PyTorch: An Imperative Style, High-Performance Deep Learning Library,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 32, 2019, pp. 8024–8035.

[5] W. McKinney, “Data Structures for Statistical Computing in Python,” in Proc. 9th Python in Science Conference (SciPy), 2010, pp. 51–56.

[6] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[7] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly Detection: A Survey,” ACM Computing Surveys, vol. 41, no. 3, pp. 1–58, Jul. 2009.

[8] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

[9] D. Jurafsky and J. H. Martin, Speech and Language Processing, 3rd ed. Upper Saddle River, NJ, USA: Pearson.

[10] C. C. Aggarwal, Data Mining: The Textbook. Cham, Switzerland: Springer, 2015.

[11] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Burlington, MA, USA: Morgan Kaufmann, 2011.