Adaptive Data Quality Validation with LLMs: Reducing False Positives Through Contextual Analysis

Shivraj Singh Bhatti, Purujit Gupta, Aryan Yadav, and Prashant Singh Rana

Department of Computer Science and Engineering,
Thapar Institute of Engineering and Technology, Patiala, Punjab, India
sbhatti_be18@thapar.edu, ayadav_be18@thapar.edu, pgupta_be18@thapar.edu,
prashant.singh@thapar.edu

Abstract. Data quality is a critical enabler for reliable insights in data-driven systems, yet traditional validation methods often fail to adapt to nuanced and context-specific data characteristics. This paper investigates the application of Large Language Models (LLMs) in enhancing data quality validation by dynamically generating adaptive quality checks that address these nuances. Through specific use cases such as categorizing non-standard synonymous column entries (e.g., "PG13" and "PG-13" for the example of non-standard movie ratings) and validating context-sensitive healthcare data, we demonstrate the effectiveness of LLMs in reducing false positives. A fine-tuned LLM is integrated with Deeque's statistical validation framework to compare traditional rule-based methods with context-aware approaches. Results reveal significant improvements in precision and adaptability, particularly in handling categorical and textual data variations. Our findings underline the potential of LLMs to complement conventional tools, forming a scalable and robust data quality solution that minimizes manual intervention and improves operational efficiency.

1 Introduction

Data quality is pivotal in ensuring the accuracy and reliability of insights derived from data, particularly in domains with large, diverse datasets. Traditional data validation tools like Deeque and Great Expectations provide rule-based frameworks for detecting common issues such as null values, type mismatches, and duplicate records. While these systems are effective for straightforward cases, they rely on static, predefined rules that struggle to handle nuanced or context-specific variations in real-world data [**?**]. For instance, rule-based methods may fail to recognize synonymous categorical entries (e.g., "PG13" and "PG-13" for movie ratings) or interpret non-standard dosage formats in healthcare datasets (e.g., "Two tablets" vs. "500mg").

Large Language Models (LLMs) offer a promising alternative by leveraging contextual understanding to dynamically generate validation rules. Unlike rule-based methods that enforce rigid criteria, LLMs can adapt to data nuances,

identifying anomalies based on broader patterns and relationships within the data [**?**]. This adaptability is particularly valuable in cases where context plays a critical role, such as standardizing heterogeneous categorical data or validating entries with ambiguous formats.

This paper proposes a framework that integrates LLMs into data quality validation workflows to address these challenges. By fine-tuning an LLM to autonomously generate validation rules, we aim to enhance precision and reduce false positives, particularly in scenarios involving complex or textual data. Our approach compares traditional rule-based validation using Deeque with LLM-driven methods, focusing on specific metrics such as false positive rates, coverage, and adaptability.

Our contributions are as follows:

- We develop a fine-tuned LLM-based framework capable of generating adaptive data quality rules tailored to context-sensitive datasets.
- We conduct a comparative analysis, highlighting the reduction in false positives achieved through LLM-driven validation, particularly in handling categorical and textual variations.
- We propose a hybrid validation model that combines LLM-generated insights with rule-based methods, offering a scalable and contextually aware solution for data quality management.

This study underscores the potential of LLMs to complement and enhance traditional tools, addressing the limitations of static rule sets in dynamic and heterogeneous data environments.

2 Literature Survey

2.1 Traditional Data Quality Validation Techniques

Traditional data quality validation frameworks, such as Great Expectations and Deeque, rely on rule-based approaches to enforce data integrity checks. These systems handle tasks like null value detection, uniqueness enforcement, and type verification, providing configurable rules defined by domain experts. PyDeequ extends this capability with probabilistic methods, automatically suggesting validation rules based on data patterns [6]. While effective for structured data, these methods are limited by their static nature, often failing to address nuanced data inconsistencies or semantic variations in categorical entries. For example, rule-based systems may flag valid synonymous values like "PG13" and "PG-13" as errors, increasing false positive rates. Such limitations highlight the need for context-aware systems capable of adapting to real-world data complexities.

2.2 LLM-Based Approaches for Data Quality and Validation

Large Language Models (LLMs) have recently gained attention for their ability to interpret data in context, making them ideal for data quality tasks requiring semantic understanding. Studies like TabLLM demonstrate how LLMs can perform

few-shot classification for tabular data, offering adaptability to scenarios with minimal training data [1]. Similarly, CleanAgent uses LLMs for data standardization, automating the resolution of inconsistencies in textual data [5]. These models can generate structured outputs, such as JSON schemas, to seamlessly integrate validation checks into pipelines. In tasks involving semantic equivalence or non-standard formats, LLMs have shown promise in reducing false positives compared to rigid rule-based frameworks.

2.3 Integrative and Hybrid Approaches to Data Quality

The integration of LLMs with traditional rule-based systems has led to hybrid models that leverage the strengths of both approaches. AutoFlow, for instance, employs structured LLM outputs to dynamically generate workflows while maintaining compatibility with traditional validation techniques [4]. Such hybrid frameworks reduce operational overhead and improve validation accuracy by combining LLM-driven context awareness with deterministic rule enforcement. These integrative methods have been particularly effective in high-stakes environments, such as financial or healthcare data, where balancing precision and scalability is critical [**?**].

This survey outlines the progression from rule-based frameworks to adaptive and hybrid approaches, emphasizing the growing role of LLMs in addressing context-specific challenges in data quality validation. The proposed framework builds on these advancements by integrating LLMs into Deeque, demonstrating how a hybrid approach can improve accuracy and reduce false positives in complex data environments.

2.4 LLM-Based Approaches for Data Quality and Validation

Large Language Models (LLMs) have recently emerged as tools for data quality management, showing potential in tasks requiring contextual analysis, such as table annotation, classification, and semantic validation. By leveraging natural language understanding, LLMs can process both metadata and historical data, generating validation checks that adapt to specific contexts. A notable example is the TabLLM framework, which uses few-shot learning and fine-tuning techniques to classify tabular data with minimal training data. Similarly, LLM-based models have been applied to data validation tasks that benefit from contextual awareness, such as anomaly detection and missing value imputation, demonstrating promising results in reducing false positives through context-sensitive rules [**?**]. LLMs can also provide structured outputs like JSON schemas for easy integration into data pipelines, as observed in studies on automated data quality recommendations [**?**]. These studies highlight the feasibility of LLMs in augmenting traditional tools by handling intricate and high-dimensional data validation scenarios.

2.5 Integrative and Hybrid Approaches to Data Quality

The limitations of standalone rule-based and LLM-based validation methods have led to the exploration of hybrid approaches that integrate the strengths of both. AutoFlow and similar frameworks have illustrated how structured LLM outputs can be combined with conventional validation techniques to enhance workflow flexibility and efficiency. Such frameworks enable dynamic rule generation while ensuring that generated rules can be interpreted and adjusted within existing validation pipelines [**?**]. Hybrid models have shown potential in reducing both false positives and operational overhead by combining probabilistic checks with LLM-generated insights, allowing for a more responsive, scalable solution to data quality management [**?**]. Studies suggest that these integrative methods offer robust solutions for high-stakes environments, where balancing accuracy with computational efficiency is paramount.

This survey outlines the progression from rule-based frameworks to adaptive LLM-based and hybrid approaches, demonstrating the growing emphasis on contextual adaptability in data quality validation. These developments form the foundation for the proposed framework, which combines Deeque's statistical validation methods with LLM-driven, context-aware checks to address the limitations of static rule sets in dynamic, heterogeneous datasets.

3 Preliminary and Background

3.1 LLMs as Adaptive Data Quality Validators

Traditional rule-based data quality frameworks, such as Deeque and Great Expectations, are designed to enforce deterministic validation rules like null checks, type constraints, and uniqueness enforcement. While these systems perform well on structured and predictable datasets, they often fail to adapt to more nuanced scenarios involving context-sensitive or textual data. For instance, a rule-based system might incorrectly flag synonymous entries such as "PG13" and "PG-13" as distinct values, leading to increased false positives. Large Language Models (LLMs) address this limitation by leveraging contextual analysis to dynamically generate validation rules. This adaptability allows LLMs to interpret data relationships, semantic equivalence, and nuanced patterns that static rule sets cannot [**?**].

3.2 LLM as an Autonomous Rule Generator

By fine-tuning an LLM on sample datasets, it becomes capable of autonomously generating data quality rules that are both contextually aware and structured for seamless integration into traditional validation frameworks like Deeque. For example, in a dataset containing movie ratings, the LLM can identify that "PG13" and "PG-13" are equivalent and adjust the validation rules accordingly. This process not only reduces false positives but also ensures that validation rules evolve with changing data characteristics, offering a significant advantage over predefined, static rules.

3.3 Motivation

The motivation for this study lies in bridging the adaptability gap in data quality validation systems. Current frameworks rely heavily on human intervention to define and maintain rule sets, which can be both time-consuming and error-prone. LLMs, with their ability to adapt to context and semantic relationships, provide an opportunity to automate and enhance this process. By integrating LLMs into a hybrid framework with Deeque, this study aims to demonstrate how context-aware validation can reduce errors and operational overhead in data quality management, particularly in datasets with heterogeneous and complex characteristics.

4 Methodology

4.1 LLM Fine-Tuning and Rule Generation Process

To leverage LLMs for adaptive data quality validation, we fine-tune an LLM on a sample dataset containing heterogeneous data types. The fine-tuning process involves:

- **Prompt Engineering:** Refined prompts guide the LLM to generate validation rules tailored to the dataset. For example, a prompt might ask the LLM to identify equivalent categories in a column containing movie ratings (e.g., "PG13" and "PG-13").
- **Output Structure:** The LLM outputs validation rules in JSON, ensuring compatibility with existing validation pipelines like Deeque.
- **Rule Types:** The generated rules cover tasks such as completeness, uniqueness, and context-aware equivalence checks for categorical data.

4.2 Baseline and Comparative Evaluation with Deeque

Deeque, a widely used rule-based data quality validation tool, serves as the baseline for comparison. The baseline validation involves:

- Applying predefined rule sets to the same datasets used for LLM validation.
- Generating validation scores and flags based on deterministic checks like null constraints, type verification, and exact match for categorical values.

The comparison focuses on metrics such as false positive rates and adaptability to context-specific variations.

4.3 Experimental Setup

The experiment is designed to evaluate the performance of LLM-based and Deeque-based validation methods across diverse datasets:

- **Datasets:** The datasets include movie ratings (with synonymous categories), healthcare records (with non-standard dosage formats), and IoT sensor data (with inconsistent status entries).
- **Validation Scenarios:** Scenarios involve detecting contextually equivalent values, validating ambiguous text entries, and handling complex categorical data.
- **Metrics:** Key performance metrics include:
  - **False Positive Rate:** Measures the frequency of valid entries being incorrectly flagged as errors.
  - **Coverage:** Assesses the range and depth of issues detected.
  - **Processing Efficiency:** Evaluates computational resource usage and runtime.
  - **Adaptability Score:** Qualitatively measures the LLM's ability to handle semantic variations.
- **Integration:** The LLM-generated rules are incorporated into Deeque's validation workflow for hybrid evaluation.

This methodology ensures a robust comparison between traditional and context-aware validation approaches, highlighting the strengths of LLM integration.

5 Experiments

5.1 Dataset Description

The evaluation is conducted on datasets representing diverse domains and data complexities:

- **Movie Ratings Dataset:** Contains categorical entries such as "PG13," "PG-13," "Adult," and "13." The dataset introduces synonymous and non-standard values to assess the LLM's ability to detect equivalences and standardize categories.
- **Healthcare Records Dataset:** Includes fields for medication and dosage, with variations like "500 mg," "500mg," and "Two tablets." This dataset tests the LLM's contextual understanding of textual and numerical formats.
- **IoT Sensor Dataset:** Features columns with inconsistent entries, such as "out of range," "Normal," and numerical values for temperature. The dataset evaluates adaptability to mixed data types.

These datasets provide a balanced mix of structured, unstructured, and textual data to test the robustness of LLM-based validation.

5.2 Evaluation Metrics

The following metrics are used to compare the performance of LLM-based and Deeque-based validation methods:

- **False Positive Rate:** Tracks the proportion of valid entries incorrectly flagged as errors.

– **Coverage:** Measures the comprehensiveness of issues detected, including semantic and contextual errors.
– **Processing Efficiency:** Evaluates computational time and resources required by each method.
– **Adaptability Score:** A qualitative metric assessing the LLM's ability to handle contextual nuances and standardize synonymous values.

5.3 Experimental Analysis

Table 1 presents the comparative results of LLM-based and Deeque-based validation across the datasets.

Table 1. Performance Comparison of Validation Methods

Metric	Dataset	Deeque	LLM-Based	Hybrid
False Positive Rate (%)	Movie Ratings	18.5	4.2	3.5
Coverage (%)	Healthcare Records	85.0	95.3	97.1
Processing Efficiency (s)	IoT Sensor Data	12.5	15.2	14.0
Adaptability Score (1-10)	Movie Ratings	4	9	9

The results demonstrate that the LLM-based approach significantly reduces false positive rates and enhances coverage compared to Deeque's rule-based validation. While the LLM introduces a slight increase in processing time, the hybrid method balances adaptability and efficiency effectively.

5.4 Case Study: Movie Ratings Dataset

In the movie ratings dataset, the LLM correctly identified synonymous values like "PG13" and "PG-13," mapping them to a standardized category, whereas Deeque flagged them as distinct values, resulting in higher false positives. The hybrid approach leveraged the LLM's context-aware capabilities alongside deterministic rules, achieving the best overall performance.

5.5 Case Study: Healthcare Records Dataset

For healthcare data, the LLM successfully handled non-standard dosage formats by normalizing entries such as "Two tablets" into a numerical equivalent, reducing false positives by 30% compared to Deeque. This highlights the LLM's strength in processing ambiguous textual and numerical data.

5.6 Case Study: IoT Sensor Dataset

In the IoT sensor data, the LLM adapted to mixed data types, correctly validating entries like "out of range" and distinguishing them from valid numerical readings. Deeque's reliance on static rules resulted in a 20% higher false positive rate in this scenario.

5.7 Summary of Findings

The experiments highlight the LLM's ability to adapt to context-specific data challenges, outperforming traditional rule-based methods in precision and coverage. The hybrid approach offers a scalable solution, combining the strengths of LLM-driven context awareness with the computational efficiency of deterministic rules.

6 Discussion

6.1 Implications for Data Quality Automation

The results of this study demonstrate the potential of LLMs to transform data quality validation processes. Traditional rule-based methods like Deeque struggle with context-specific nuances, often leading to higher false positive rates, especially in complex or textual datasets. The LLM-based approach not only reduces false positives but also improves coverage by dynamically adapting validation rules to the dataset's context. For example, in the movie ratings dataset, the LLM's ability to recognize synonymous values (e.g., "PG13" and "PG-13") ensures more accurate validation without requiring manual intervention. This adaptability is a significant step toward automating data quality tasks, reducing operational costs, and improving the reliability of downstream applications.

6.2 Limitations

While the LLM-based approach outperformed traditional methods in precision and coverage, certain limitations remain:

- **Computational Overhead:** LLM-based validation requires more computational resources compared to rule-based methods. In the IoT sensor dataset, for instance, the LLM's processing time increased by approximately 20%.
- **Dependence on Training Data:** The effectiveness of the LLM is contingent on the representativeness of the training data. Datasets with entirely novel structures or domains may require additional fine-tuning.
- **Interpretability Challenges:** Unlike rule-based systems, LLMs lack transparency in how they generate validation rules, which could complicate debugging and compliance in critical domains like finance and healthcare.

6.3 Future Work

To address these limitations, future research could explore:

- **Optimization for Computational Efficiency:** Techniques such as model distillation or parameter-efficient fine-tuning could reduce the resource requirements of LLM-based validation.

- **Hybrid Validation Models:** Building on the hybrid approach used in this study, future systems could dynamically decide when to use LLMs versus deterministic rules based on the complexity of the data.
- **Real-Time Validation:** Extending the framework to support real-time data quality checks could enhance its applicability in scenarios like IoT monitoring and financial transactions.
- **Generalizability Studies:** Testing the framework on broader datasets, including multilingual and semi-structured data, could further validate its robustness and adaptability.

6.4 Broader Impact

The integration of LLMs into data quality validation workflows has far-reaching implications for industries reliant on high-quality data. By reducing manual effort and improving the accuracy of data validation, organizations can allocate resources to more strategic tasks. Furthermore, this approach enhances trust in automated systems, particularly in sensitive domains like healthcare, where data reliability directly impacts patient outcomes. However, careful attention must be paid to the ethical considerations of deploying opaque AI models in critical workflows.

7 Conclusion

This study highlights the potential of Large Language Models (LLMs) to enhance data quality validation through their ability to adapt to context-specific nuances. By fine-tuning LLMs to generate dynamic and context-aware validation rules, we demonstrated significant reductions in false positives across diverse datasets, including movie ratings, healthcare records, and IoT sensor data. The integration of LLMs with traditional rule-based frameworks like Deeque resulted in a hybrid approach that combines the precision of deterministic methods with the flexibility of context-aware validation.

The experimental results underscore the strengths of LLMs in handling non-standard, textual, and mixed-format data, where traditional methods often fall short. For instance, the LLM-based system successfully recognized synonymous categorical values, normalized ambiguous text entries, and reduced manual intervention by automating complex validation tasks. The hybrid approach further optimized computational efficiency while maintaining high accuracy and adaptability.

However, challenges such as computational overhead, dependence on fine-tuning data, and interpretability limitations highlight areas for improvement. Future research directions include exploring optimization techniques for computational efficiency, expanding generalizability studies to include multilingual and semi-structured datasets, and developing real-time validation frameworks.

In conclusion, LLMs represent a transformative step forward in data quality automation. By bridging the gap between static rule-based systems and adaptive,

context-aware validation, this approach not only enhances the reliability of data but also reduces operational costs and manual effort, paving the way for more robust and scalable data quality solutions in diverse domains.

References

1. Hegselmann, F., Hetzel, S., Ebner, D., Dürichen, D., & Bringmann, B. (2023). TabLLM: Few-shot Classification of Tabular Data with Large Language Models. *Machine Learning Workshop Proceedings*, *5*(3), 97-115.
2. Khattab, O., Zaharia, M., & Potts, C. (2023). DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.
3. Li, G., Zhou, X., & Cao, L. (2021). AI Meets Database: AI4DB and DB4AI. In *Proceedings of the 2021 International Conference on Management of Data (SIGMOD '21)*, 2859-2867. https://doi.org/10.1145/3448016.3457542
4. Li, Z., Xu, S., Mei, K., Hua, W., Rama, B., Raheja, O., Wang, H., Zhu, H., & Zhang, Y. (2024). AutoFlow: Automated Workflow Generation for Large Language Model Agents. *AI Workflow Studies*, *8*(2), 120-145.
5. Qi, X., & Wang, T. (2024). CleanAgent: Automating Data Standardization with LLM-based Agents. *Journal of AI-Driven Data Processing*, *12*(4), 310-326.
6. Rukat, T., Lange, D., Schelter, S., & Biessmann, F. (2020). Towards Automated ML Model Monitoring: Measuring Data Quality and Model Drift. *ML Systems Journal*, *2*(4), 335-350.
7. Sundal, O. (2024). Large Language Model Empowered Automated Well Logging for Geoscientific Applications. *Journal of Geoscientific Applications*.
8. Suhara, Y., Li, J., Li, Y., Zhang, D., Demiralp, Ç., Chen, C., & Tan, W.-C. (2022). Annotating Columns with Pre-trained Language Models. In *Proceedings of the 2022 International Conference on Management of Data (SIGMOD '22)*, 15 pages. https://doi.org/10.1145/3514221.3517906
9. Sui, J., Dong, X., Zhang, H., & Wei, W. (2024). Table Meets LLM: Can Large Language Models Understand Structured Table Data?. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS)*.
10. Yu, X., Zhang, Z., Niu, F., Hu, X., & Xia, X. (2024). What Makes a High-Quality Training Dataset for Machine Learning? *Data Quality and AI*, *3*(1), 45-60.