Know what you don't know - Leveraging LLM domain expertise

Ruben Chiche
rubenchiche0317@gmail.com

Ophir Haroche
ophirbh@gmail.com

1 Abstract

In this paper, we introduce an automated data validation framework that leverages Large Language Models (LLMs) to enhance Exploratory Data Analysis (EDA). Our approach systematically evaluates datasets by identifying missing values, feature distribution anomalies, and inconsistencies, providing actionable insights with minimal human intervention. By simulating domain expertise, LLMs improve the accuracy, efficiency, and reproducibility of data validation, addressing limitations of traditional expert-driven approaches such as human error, subjectivity, and time constraints. We assess our framework's effectiveness by benchmarking its performance against expert-driven validation across diverse datasets. Our results demonstrate that LLMs can significantly enhance data validation workflows, serving as valuable tools to complement human expertise and mitigate common pitfalls in manual EDA.

2 Problem Description

Exploratory Data Analysis (EDA) is a crucial step in the data science pipeline, often dictating the success of downstream tasks such as model development and deployment. A key aspect of EDA is data validation, which involves identifying and addressing potential issues within a dataset—such as missing values, imbalanced distributions, or data leakage risks. Traditionally, this process is a collaborative effort between data scientists and domain experts, who jointly inspect and interpret data characteristics based on their respective expertise. However, this collaboration is prone to several limitations, including human error, subjective judgment, communication gaps, and time-consuming iterations. These challenges can result in incomplete or suboptimal validation of datasets, which may ultimately compromise the reliability and performance of the developed models. As datasets become larger and more complex, and project timelines tighten, there is a growing need for automated, scalable, and robust methods to enhance the data validation process. A more systematic approach that can consistently identify a broader spectrum of data quality issues, while reducing the reliance on manual interpretation and subjective decision-making, would significantly improve the efficiency and accuracy of EDA.

3 Solution Overview

We propose an automated data validation framework that leverages the domain knowledge capabilities of Large Language Models (LLMs) to address common pitfalls in EDA. Given a dataset and a problem statement, we utilize LLMs to simulate the role of a domain expert, systematically evaluating data quality issues and providing actionable insights.

The proposed framework comprises several key components:

1. **Missing Features Evaluator**: Detects essential features that are absent from the dataset based on the context of the given problem statement.

2. **Missing Categorical Values Evaluator**: Identifies missing or rare categorical values that might indicate incomplete data collection or potential data leakage.

3. **Categorical Features Distribution Evaluator**: Analyzes the distribution of categorical variables to highlight imbalances, anomalies, or unexpected patterns that could affect model performance or introduce bias.

4. **Numerical Features Evaluator**: Examines the distribution and missingness in numerical features to detect data quality issues such as skewness, inconsistencies, or suspicious trends.

By harnessing the reasoning and general knowledge capabilities of LLMs, this approach reduces dependence on manual inspection and expert-driven validation. The system enhances accuracy, reproducibility, and speed in the EDA phase, while mitigating risks associated with human error and subjective interpretation. Ultimately, this framework aims to streamline the transition from raw data to reliable, model-ready datasets.

4 Experimental Evaluation

4.1 Evaluation Process

To evaluate our automated data validation framework, we compare its performance against human domain experts. The evaluation follows these steps:

1. Domain experts manually examine datasets to identify data quality issues, focusing on missing features and missing categorical values, as we learned in test that distributional and numerical characteristics are harder for them to assess.
2. Our framework independently analyzes the same datasets using LLMs to flag potential issues.
3. We compare outputs from experts and the framework and have them reviewed by the experts to establish ground truth (GT) labels for missing features and categorical values.
4. Since numerical and distributional characteristics are assumed valid, we introduce a perturbation test: altering a numerical feature at a time to test if our system detects the change and deem it invalid.

4.2 Evaluation Metrics

To quantify the performance of our automated system, we use the following evaluation metrics:

- **Precision**: Measures the proportion of correctly identified issues among all flagged issues. A higher precision indicates fewer false positives.
- **Recall**: Measures the proportion of actual data quality issues correctly identified by the system. A higher recall indicates fewer false negatives.
- **F2-score**: A weighted harmonic mean of precision and recall, giving more importance to recall. Since recall is more critical in data validation (i.e., missing an issue is more detrimental than a false alarm), we use F2 to prioritize recall over precision.
- **Accuracy**: Measures the proportion of correctly identified instances (both issues and non-issues) out of all instances. A higher accuracy indicates better overall performance in distinguishing issues from non-issues.

Finally, it is important to note that we have not dealt with semantics. Meaning, if the expert and the LLM articulated the same concept in different ways, we aligned it in our experiment recorr. In addition, for some of the datasets it was hard to decide on the domain expert persona, and in some cases the GT set was open for interpretation, rather than a clear cut decision.

4.3 Results

4.3.1 Missing Features Evaluation

Figure 1 presents the performance metrics for identifying missing features across different methods.

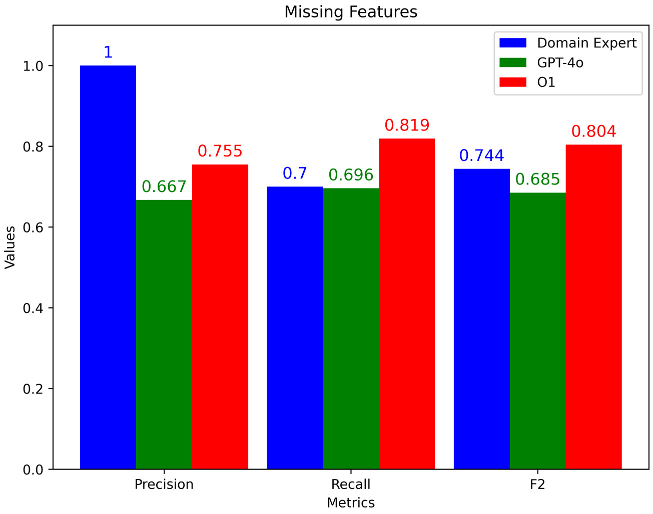

Figure 1: Comparison of Precision, Recall, and F2-score across different methods for missing features detection.

The domain expert achieved the highest **Precision (1.0)**, indicating that all detected missing features were correct. However, its **Recall (0.7)** was lower than that of O1, suggesting that some missing features were not identified. GPT-4o showed a more balanced performance, with moderate precision and recall. O1 demonstrated the highest **Recall (0.819)** and the best **F2-score (0.804)**, suggesting a strong ability to detect missing features comprehensively, even though with a slightly lower precision than the domain expert.

These results highlight that while human experts provide highly precise annotations, automated methods—especially O1—offer improved recall and may serve as valuable complements in data validation workflows.

4.3.2 Missing Categorical Values Evaluation

The domain expert achieved the highest **Precision (1.0)**, meaning all identified missing categorical values were correct. However, its **Recall (0.843)** was lower than perfect, indicating some missed cases. GPT-4o had lower precision and recall, with an **F2-score (0.545)** that reflected a more balanced performance but still lagged behind the expert. O1 had consistent but lower precision and recall values (**0.654** across all metrics), indicating a stable but less accurate performance.

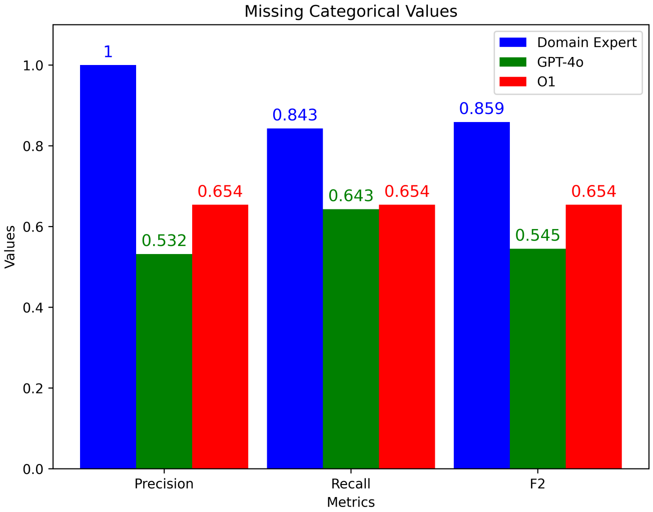

Figure 2: Comparison of Precision, Recall, and F2-score across different methods for missing categorical values detection.

4.3.3 Categorical Features Distribution Validation

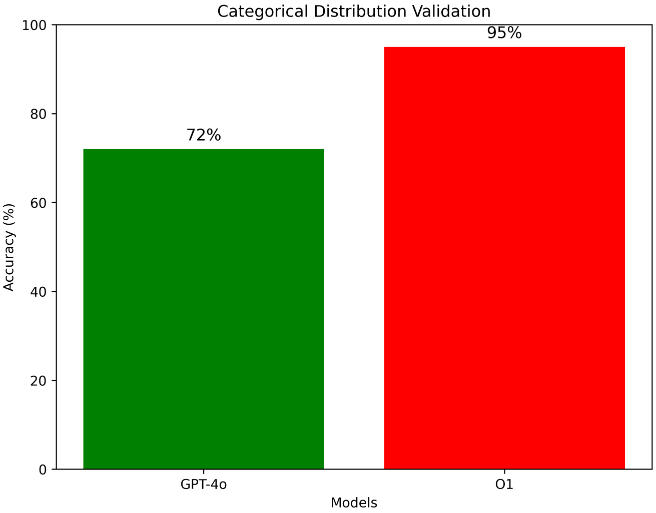

Figure 3: Accuracy comparison for categorical features distribution validation across different Models.

The categorical features distribution validation results show that O1 performed significantly better than GPT-4o, achieving an accuracy of **95%** compared to **72%**. This suggests that O1 is more reliable for detecting categorical features distribution shifts, whereas GPT-4o exhibits moderate accuracy in this task. In this test, we did not include domain expertise as a baseline because determining whether such distribution is valid proved to be too complex for our experts. These findings emphasize the need for combining automated and expert-driven approaches for optimal data validation. Additionally, the categorical feature distribution of the dataset was assumed to be valid, which may not always be the case in real-world scenarios.

4.3.4 Numerical Features Validation Using Descriptive Statistics - Baseline

To evaluate how well different models validate numerical features distribution statistics, we conducted a baseline validation experiment. In this experiment, we assessed the accuracy of GPT-4o and O1 in identifying whether numerical features distribution statistics in a given dataset aligned with expected patterns.

Figure 4 presents the accuracy scores for both methods. O1 achieved a slightly higher accuracy (90%) compared to GPT-4o (89%), indicating that both models performed well in detecting distribution statistics conformity. The marginal difference suggests that O1 may have a slight advantage in this assessment, but both models are highly effective.

These results indicate that LLMs can serve as reliable tools for validating numerical distribution statistics, reducing the need for extensive manual inspection.

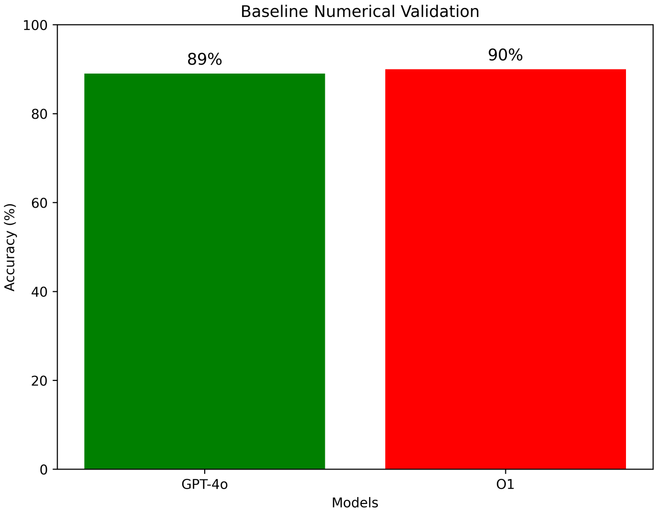

Figure 4: Accuracy comparison for numerical distribution baseline validation across different methods.

4.3.5 Numerical Features Validation Using Descriptive Statistics - Perturbed

To further test the robustness of our models, we introduced perturbations to numerical distribution statistics and evaluated how well GPT-4o and O1 detected these changes. This experiment aimed to assess the sensitivity of each model to deviations from expected numerical patterns.

As shown in Figure 5, both GPT-4o and O1 achieved an accuracy of 94%, demonstrating strong performance in detecting such shifts. These results suggest that both models are capable of identifying perturbations with high precision, making them effective tools for automated numerical distribution validation.

The increased accuracy in this experiment compared to the baseline suggests that both models are more responsive to deviations from expected distributions.

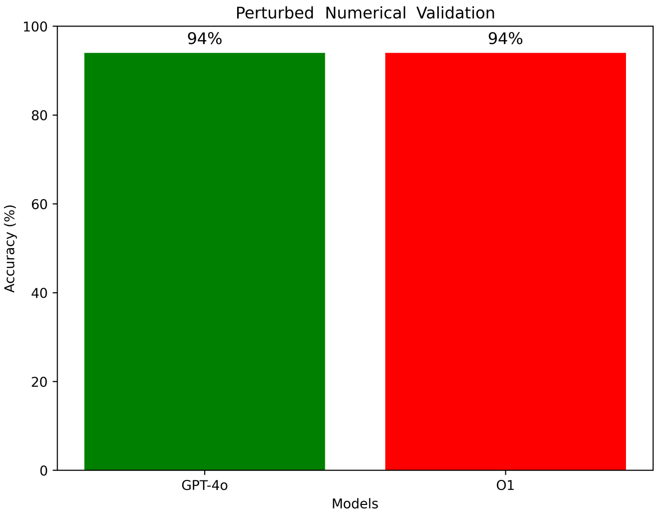

Figure 5: Accuracy comparison for numerical distribution baseline validation across different methods.

5 Related Work

The integration of Large Language Models (LLMs) into data validation and Exploratory Data Analysis (EDA) is an emerging research area with limited prior work, thus we could not find a solution scheme that is somewhat close to ours to enable direct comparison, therefore we compared to the traditional human domain-expert method. While comprehensive studies directly addressing automated data validation using LLMs are scarce, a few notable works provide foundational insights into this intersection.

Alexander et al. (2024) [1] investigated the capability of LLMs to generate data validation tests. Their study assessed various prompting strategies, learning modes, and temperature settings to evaluate the quality and consistency of LLM-generated validation tests. The findings suggest that while LLMs can complement traditional data validation approaches, their outputs require careful evaluation by experienced analysts to ensure reliability.

In a practical application, Mirateterekhova (2025) [2] explored the automation of data quality checks using LLMs within a data warehouse context. The proposed approach leverages LLMs to generate tests automatically for various data sources, significantly reducing the time required for data validation from several hours to mere minutes. This work demonstrates the potential of LLMs to streamline data quality assurance processes in real-world settings.

Additionally, InsightPilot, an LLM-based automated data exploration system, has been introduced to simplify the data exploration process [3]. InsightPilot collaborates with LLMs to issue a sequence of analysis actions, explore data, and generate insights based on natural language queries, thereby enhancing the efficiency of data exploration tasks.

In these contributions, one element that we missed and that appears on our project was the problem statement. A validity of the a dataset is also derived from the problem at hand. For example, a dataset that aims to predict diamond prices and miss the feature ”natural/lab-grown” is not valid for all kinds of diamonds but is valid to each one of these types alone.

6 Conclusion

Our study demonstrated that LLMs can effectively enhance data validation in EDA, given a problem statement, by identifying missing features, values, distribution anomalies, and inconsistencies. We found that while human experts achieve higher precision, LLMs may provide superior recall, detecting a broader range of issues. Not to mention additional benefits such as shorter process times. One interesting finding was that when the LLM suggests an addition to the dataset or points to a phenomena, even if that addition is not mandatory or that phenomena does not deem the feature as invalid per the LLMs judgement, it seems (at least on the surface) that it may provide added value. One downside of LLMs that we implicitly tested was consistency, and according to our findings it seems that O1 and the new line of reasoning models show improvement on that area. The best results were obtained by combining both approaches, leveraging human expertise for accuracy and LLMs for comprehensive issue detection. Additionally, our experiments showed that LLMs are particularly effective in categorical and numerical distribution validation, though they sometimes generate false positives (in such cases, sometimes their reasoning outputs mention the issues even though the answer is false). These findings highlight the potential of LLMs to improve data validation workflows by reducing manual effort, and mitigating common human errors.

References

[1] Alexander et al. ”LLMs for Data Validation Testing.” *arXiv*, 2024. Available: https://arxiv.org/pdf/2310.01402

[2] Mirateremkhova, M. ”How LLM Can Validate Data.” *Medium*, 2025. Available: https://medium.com/@mirateremkhova/how-llm-can-validate-data-e61c8ada4fa0

[3] Microsoft Research. ”InsightPilot: An LLM-Empowered Automated Data Exploration System.” 2023. Available: https://www.microsoft.com/en-us/research/publication/insightpilot-an-llm-empowered-automated-data-exploration-system/