Enhancing Data Quality through Generative AI: An Empirical Study with Data

Pan Singh Dhoni

1Five Below

December 7, 2023

Abstract

In today's increasingly data-driven landscape, organizations are shifting their focus toward leveraging data analytics for strategic decision-making. As data becomes a cornerstone of operational and strategic activities, the quality of this data has emerged as a non-negotiable aspect for organizations. Lack of attention to data quality can not only result in considerable revenue losses but can also cripple the effectiveness of analytics, causing misinformed decisions and strategic errors. Against this backdrop, this empirical study delves into the innovative avenue of utilizing Generative Artificial Intelligence (AI) as a mechanism for enhancing data quality.

The research aims to explore multiple facets of organizational operations ranging from technical infrastructure to business strategy to ascertain the potential advantages offered by Generative AI. Utilizing a mix of qualitative and quantitative methods, we conducted in-depth interviews, case studies, and simulations to evaluate the impact of Generative AI on data quality.

Our findings reveal a multi-layered benefit structure. Notably, we found that Generative AI is not a replacement for existing, traditional methods of data quality assurance but serves as a powerful supplement. It augments traditional methods by increasing the accuracy of data, thereby offering a more reliable foundation for analytics. Additionally, the use of Generative AI can streamline workflows, enhancing productivity among various roles including solution architects and software developers. Moreover, it facilitates a more nuanced and accurate requirement gathering process, enabling businesses to fine-tune their data analytics strategies more effectively.

In conclusion, our study establishes that integrating Generative AI into data quality management processes can not only resolve immediate issues surrounding data accuracy but also lead to long-term organizational benefits, such as higher efficiency and more effective decision-making. This research serves as a pioneering step in the intersection of Generative AI and data quality, setting the stage for future studies and real-world applications.
Enhancing Data Quality through Generative AI: An Empirical Study with Data

Pan Singh Dhoni (ORCID: 0009-0006-7143-5353)

Abstract— In today's increasingly data-driven landscape, organizations are shifting their focus toward leveraging data analytics for strategic decision-making. As data becomes a cornerstone of operational and strategic activities, the quality of this data has emerged as a non-negotiable aspect for organizations. Lack of attention to data quality can not only result in considerable revenue losses but can also cripple the effectiveness of analytics, causing misinformed decisions and strategic errors. Against this backdrop, this empirical study delves into the innovative avenue of utilizing Generative Artificial Intelligence (AI) as a mechanism for enhancing data quality.

The research aims to explore multiple facets of organizational operations—ranging from technical infrastructure to business strategy—to ascertain the potential advantages offered by Generative AI. Utilizing a mix of qualitative and quantitative methods, we conducted in-depth interviews, case studies, and simulations to evaluate the impact of Generative AI on data quality.

Our findings reveal a multi-layered benefit structure. Notably, we found that Generative AI is not a replacement for existing, traditional methods of data quality assurance but serves as a powerful supplement. It augments traditional methods by increasing the accuracy of data, thereby offering a more reliable foundation for analytics. Additionally, the use of Generative AI can streamline workflows, enhancing productivity among various roles including solution architects and software developers. Moreover, it facilitates a more nuanced and accurate requirement gathering process, enabling businesses to fine-tune their data analytics strategies more effectively.

In conclusion, our study establishes that integrating Generative AI into data quality management processes can not only resolve immediate issues surrounding data accuracy but also lead to long-term organizational benefits, such as higher efficiency and more effective decision-making. This research serves as a pioneering step in the intersection of Generative AI and data quality, setting the stage for future studies and real-world applications.

Index Terms— accuracy, ai, dq, data quality, chatgpt, data, generative ai, genai, information technology, synthetic, timeliness, uniqueness, validation

I. INTRODUCTION

In a world increasingly governed by data-driven decision-making, the importance of maintaining high data quality has never been more pronounced. According to a report by Gartner, organizations face an annual cost of approximately $9.7 million due to poor data quality. Similarly, IBM estimates that suboptimal data quality results in annual losses amounting to a staggering $3.1 trillion for U.S. businesses alone. These staggering figures highlight the urgency for organizations to reevaluate and upgrade their data management strategies.

One such avenue for innovation is the application of Generative Artificial Intelligence (AI). While the intersection of Generative AI and data quality has not been widely explored, preliminary research and applications in related fields suggest that AI has the potential to offer transformative solutions for data quality challenges. Could Generative AI serve as a valuable tool in enhancing data quality? This is the central question that this empirical study aims to answer.

Specifically, this study will:
- Investigate the roles that Generative AI can play in enhancing various dimensions of data quality such as Consistency, Uniqueness, and Completeness.
- Evaluate the impact of Generative AI technologies on organizational workflows, focusing on roles like solution architects and developers.
- Examine future trends, including the integration of Large Language Models (LLMs) in data quality tools.

It is important to note that our research operates under the assumption that organizations are not providing their data to Generative AI companies for model training, thereby addressing potential ethical concerns related to data privacy.

By delving into the unexplored territory of Generative AI's role in enhancing data quality, this study aims to offer new insights and practical guidelines for organizations looking to fortify their data management strategies, thereby potentially saving millions of dollars and enhancing decision-making capabilities.
II. WHAT IS ARTIFICIAL INTELLIGENCE (AI)?

Artificial intelligence (AI) [5] is intelligence—perceiving, synthesizing, and inferring information—demonstrated by machines, as opposed to intelligence displayed by humans or by other animals.

![Artificial Intelligence Diagram](image)

Machine learning is a subset of AI, and deep learning (Natural Language processing (NLP), Convolutional Neural Network (CNN), and Recurrent Neural Network (RNN)) is a subset of machine learning (See Fig 1). Deep and machine learning have two important types: supervised learning (labelled data) and unsupervised learning.

III. GENERATIVE AI

As per Wikipedia [6], Generative artificial intelligence (AI) is artificial intelligence capable of generating text, images, or other media, using generative models.

Training generative AI [7] involves a multi-step process that includes preparing data, fine-tuning, and optimization. The goal is to enable the AI system to understand language patterns and generate coherent and contextually relevant text. Let's break down the process in detail:

**Data Collection and Preprocessing:** Collect a diverse and representative dataset that aligns with the type of content you want the AI to generate. For text generation, this could be a collection of books, articles, websites, etc. Clean and preprocess the data by removing irrelevant or noisy content.

**Model Architecture Selection:** Choose an appropriate architecture for your generative AI model. GPT-3 and similar models use transformer architectures. You might need to decide on the model size, number of layers, and other hyperparameters based on your computational resources and the complexity of the tasks you want the AI to perform.

**Tokenization:** Tokenization involves breaking down the input text into smaller units called tokens. Tokens can be as short as a single character or as long as a word. This step is crucial for feeding text into the model. Each token is usually associated with an embedding vector that the model uses for processing.

**Model Training:** Training a large generative AI model is computationally intensive and usually requires specialized hardware like GPUs or TPUs. The training process involves presenting the model with input sequences and having it predict the next token in the sequence. This process is known as "unsupervised learning" as the model learns patterns and relationships from the data without explicit labels.

**Loss Function:** During training, a loss function is used to measure the difference between the predicted tokens and the actual tokens. This guides the model to adjust its internal parameters to improve its predictions.

**Backpropagation and Optimization:** Backpropagation is a mathematical process that calculates how the model's parameters should be adjusted to minimize the loss. Optimization algorithms like Adam or SGD (Stochastic Gradient Descent) are used to update the model's parameters based on these calculations.

**Fine-tuning:** After the initial training, you might fine-tune the model on specific tasks or domains to improve its performance in those areas. This involves training the model on a narrower dataset related to the task you want it to perform.

**Evaluation:** Evaluating the model's performance is essential. Metrics like perplexity (how well the model predicts the next token) or specific task-related metrics can be used.

**Deployment and Inference:** Once the model is trained and evaluated, you can deploy it for generating text or performing specific tasks. Users interact with the model by providing prompts, and the model generates text based on its training.

Ethical Considerations: Large generative models can produce biased, offensive, or inappropriate content. It's important to implement safeguards, content filtering mechanisms, and ethical guidelines to ensure responsible and safe usage.

Remember that training large generative AI models requires significant computational resources, expertise in machine learning, and careful consideration of ethical implications. Additionally, it's a dynamic field, so staying updated with the latest research and best practices is crucial.

IV. DATA QUALITY

As we already talked about, bad data cost too much for each organization, that means data quality is very important for each organization.

Data quality has different dimensions, we will discuss these dimensions each one time:
**Consistency:** Data in sync with publisher and consumer, mean data flow between systems A to Z are in same. We have been not losing data between. System A is sending table A with 10 records with five columns, are we getting the same records and column in System Z.

**Timeliness:** Data arrives from system A to Z at 10 AM EST, does it is happening every day. Timeliness talks about the data available at the time it needed.

**Uniqueness:** Duplicate records will not be available, let say Steve has five records with same data set (same DOB, age, weight, height, and SSN), it is not unique. It should be one records for him.

**Validity:** Data should be valid, suppose a company sells $ 1 product, amount section can’t contain 1 million against one product or let say USA phone digit contains 10 number including area code, in case we have been getting 12 or 13 digits, its not a valid phone no.

**Accuracy:** Data must be accurate, DC should be USA capitol, not LA.

**Completeness:** Required data must be available, in case of Employee table, if EID can’t be missing.

V. METHODOLOGY

Few methods have been taken from research paper [5][7][8]

**Method 1** - Before we discuss Method 1, the underlying idea is to input instructions in natural English language. In this context, one can specify a validation or quality rule in English. Behind the scenes, this instruction is then converted into SQL to produce data quality output (As rules are mostly mathematical conditions). For this approach, we’ve employed SODA GPT [9]. You can input an English statement related to data quality, and it will generate a YAML description in English, which you can then utilize for your desired outcome.

Query to SODA GPT "I was surprised when the Order table didn't capture orders dropped from the platform. Can you make sure this doesn't happen again?"

The AI responds:

"Sure, no problem! Here's a SodaSQL check that does that for you. Copy and paste this into your checks YAML file”.

Code snippet is:

```yaml
checks:
  - order:
      schema:
        when: "required column: missing: [Order]"
```

Users can write tests using SQL queries to define what constitutes correct data. These tests can be used to catch common data issues, like missing values, duplicate rows, or values that fall outside of expected ranges.

SodaSQL uses YAML files to describe which tests should be run against which tables and columns. This provides a clear and human-readable way to configure your data tests.

Apart from tests, SodaSQL can compute various metrics on data, like row counts, distinct counts, missing values, and more.

While SodaSQL is an open-source tool that can be run locally, there's also an associated cloud service called Soda Cloud. Soda Cloud provides a more comprehensive platform with features like data observability, collaboration, and alerting. With Soda Cloud, you can get alerted in real-time when your data tests fail.

**Method 2** For this research paper, we have built Data quality framework, considering Azure toolset, data resides at Azure BLOB delta tables involves leveraging Azure native services and open-source technologies to monitor and improve the five dimensions you specified: Timeliness, Uniqueness, Validity, Accuracy, and Completeness.

1. **Ingestion Layer:** Build ingestion pipeline using Azure Data factory & Databricks, orchestrated using ingestion using Azure Data Factory or Azure Databricks, so data will arrive on time. It fulfills timeliness.

2. **Delta lake on Azure BLOB follows ACID transactions, it has time-travel features.**

3. **For Uniqueness:** During data ingestion, we have generated unique identifiers or hash codes for rows and validate for duplicates.

4. **For Validity:** With the help of generative AI, we have created validation rules such as (code will be generated using the help of Generative AI tools):

   ```python
   df = df.filter(df['age'] > 0)
   ```
5. **For Completeness:** With the help of generative AI, we have created validation rules such as:

```python
for col in df.columns:
    df = df.filter(df[col].isNotNull())
```

6. **Monitoring Layer:** Integrated with Azure Monitor and Azure Log Analytics to create custom monitoring solutions.

7. **Notification Layer:** Used Azure Logic Apps to set up email/SMS alerts if data is not ingested on time.

8. **Logging and Reports:** Log quality metrics in a separate Delta Lake table.

9. **Self-Service Dashboard:** Used Power BI to create dashboards for business users to monitor data quality metrics.

These templates and chatbot messages are just a starting point. Customizing them according to the specific metrics or characteristics of each cluster will make your marketing efforts even more effective.

**Method 3: Customized LLM for Data Quality Rules:**

```python
def check_conformity(self):
    """Logic will be here."
    Returns:
    bool: True if the data conforms to the schema, False otherwise.
    ....
    # Check for required columns
    if len(self.data.columns) < 10:
        return False
    return True

def check_accuracy(self):
    """Logic will be here."
    Returns:
    bool: True if the data is accurate, False otherwise.
    ....
    # Check for invalid values
    if self.data.apply(np.Series.is_valid).sum() < self.data.shape[0]:
        return False
    return True

def check_integrity(self):
    """Logic will be here."
    Returns:
    bool: True if the data is intact, False otherwise.
    ....
    # Check for missing values
    if self.data.isna().sum().sum() > 0:
        return False
    return True

def check_timeliness(self):
    """Logic will be here."
    Returns:
    bool: True if the data is timely, False otherwise.
    ....
    # Check for data older than a certain date
    if self.data["data"].min() < pd.Timestamp('2023-10-26'):
        return False
    return True

def check_all(self):
    """Checks all of the data quality rules."
    Returns:
    bool: True if all of the data quality rules pass, False otherwise.
    ....
    return all(
        self.check_completeness(),
        self.check_consistency(),
        self.check_integrity(),
        self.check_consistency(),
        self.check_timeliness()
    )
```

```python
import pandas as pd
class DataQualityCustomLlm:
    def __init__(self, df):
        self.data = df

    def check_completeness(self):
        """Logic will be here."
        Returns:
        bool: True if the data is complete, False otherwise.
        ....
        return self.data.isnull().sum().sum() == 0

    def check_integrity(self):
        """Logic will be here."
        Returns:
        bool: True if the data is consistent, False otherwise.
        ....
        # Check for duplicate values
        if self.data.duplicated().sum() == 0:
            return False
        # Check for data types
        if self.data.dtypes.unique().shape[0] == 0 or self.data.dtypes.iloc[0] == object:
            return False
        return True

    def check_accuracy(self):
```

**User Interaction:** The user initiates the process by providing data and a request to "check the data for completeness" or any other data quality check. The LLM Model processes the user's request.

**LLM Model:** The LLM model can parse and understand the user's prompt. This means that when the user asks for a specific data quality check (e.g., checking for completeness), the LLM model knows what to do and can potentially execute the respective functions or provide relevant responses.

**Data Quality Framework:** This framework consists of various functions to ensure the quality of the data:

- `check_completeness()`: Determines if the data is complete.
- `check_integrity()`: Checks if the data is consistent.
check_conformity(): Validates if the data conforms to a specific schema.
check_all(): A wrapper function that runs all quality checks on the data.

The results of these checks can be interpreted as either "The data quality is good" or "The data quality is bad" based on the function's outcome.

Integration: When the LLM model receives the user's prompt to check data quality, it can invoke the relevant function(s) from the Data Quality Framework.

Once the checks are performed, the LLM model can convey the results back to the user.

VI. RESULTS

As per articles [10][11] and research below are findings:

Generative AI can be a powerful tool for improving data quality across multiple dimensions. Here are some of the ways generative AI can be employed to enhance data quality:

Data Imputation and Completion

1. Completing Missing Data: Generative models like Generative Adversarial Networks (GANs) can be trained to generate realistic data points that could fill in the gaps for missing or incomplete data, even Synthetic data will be used to fill missing data into the database, this will improve data completeness.
2. Time Series Forecasting: Generative models can be used to forecast future data points in a time series, ensuring timeliness in data reporting.

Data Validation and Cleansing

1. Data Anomaly Detection: Generative models can be trained to understand what 'normal' data should look like. Any data point that deviates significantly from the generated 'normal' data could be flagged as an anomaly.
2. Data quality validation check will be performed using Generative AI.
3. Data Standardization: Generative models can produce data that adheres to specific formats or standards, which can be used to validate and correct existing data.

Data Augmentation

1. Enhancing Datasets: In scenarios where the data is imbalanced (common in classification problems), generative models can produce additional synthetic samples to balance the dataset.
2. Feature Engineering: Generative models can be used to create new features that may better represent the underlying structure of the data, thereby improving its quality.

Data Simulation and Testing

1. Realistic Test Cases: Generative AI can create synthetic yet realistic datasets that mimic actual production data. These synthetic datasets can be used for more robust testing of data pipelines and applications.
2. Stress Testing: Using generated data, you can simulate extreme but plausible scenarios to test the resilience and accuracy of your data systems.

Data Governance and Compliance

1. Data Masking: Generative AI models can produce synthetic data that retains the statistical characteristics of the original dataset but doesn't include sensitive or personally identifiable information (PII).
2. Auditing and Monitoring: Generative models can be integrated into your data pipelines to create baselines and benchmarks for data quality. Any deviation from these baselines can trigger alerts.

Benefits

1. Cost-Efficiency: The ability to generate synthetic data reduces the need for expensive data collection and labelling.
2. Accuracy: Generated data can be used to train more robust machine learning models, ultimately leading to more accurate analyses and predictions.
3. Compliance: By generating synthetic data without sensitive attributes, businesses can perform robust data analytics without violating privacy regulations like GDPR or HIPAA.
4. Agility: The rapid prototyping and testing enabled by generative AI can accelerate the data preparation and model training phases, enabling businesses to adapt more quickly to changing conditions.
5. Reliability: By augmenting existing data with high-quality, generated data, businesses can create more reliable and resilient analytics platforms.

By integrating generative AI into your data quality framework, you can significantly enhance the timeliness, uniqueness, validity, accuracy, and completeness of your data.

VII. CONCLUSION

The significance of data quality in today's technologically driven business environment cannot be overstated. As organizations increasingly look to data analytics for deriving actionable insights, the quality of the underlying data becomes a pivotal factor. The central tenet of our research focused on the role of Generative Artificial Intelligence (AI) in augmenting data quality, a subject that has received insufficient attention thus far.

Our findings indicate a multi-faceted role for Generative AI in improving data quality. Importantly, Generative AI is not envisioned as a replacement for human expertise but as a supplementary tool that amplifies the human capability to manage data quality. One key discovery was the absence of inbuilt frameworks within current Generative AI technologies for directly addressing data quality parameters like Consistency, Uniqueness, and Completeness. Nonetheless, Generative AI offers alternative routes to enhancing these
Aspects. For example, solution architects can create robust design schematics using Generative AI, which can then be specifically tailored for individual organizational needs. Moreover, developers and Subject Matter Experts (SMEs) can avail themselves of Generative AI assistance in crafting accurate validation rules, thus indirectly contributing to data quality enhancement.

The study was conducted with the ethical consideration that organizations would not be sharing their data for training Generative AI models, alleviating potential data privacy concerns.

Additionally, our research offers insights into the simplification of data processing rules. Generative AI technologies are positioned to recommend appropriate tools for dealing with massive datasets, thereby streamlining organizational efforts to maintain data quality. This is of particular significance for organizations grappling with the challenges of Big Data.

Looking to the future, we see an exciting trend wherein data quality tools may increasingly integrate Large Language Models (LLMs). This negates the need for investments in complex graphical user interfaces (GUIs) and promises more intuitive, efficient, and enhanced tools that can assist users in managing data quality.

In summary, while Generative AI may not replace existing data quality management methods, its potential to significantly augment these methods is promising. Our research serves as an initial foray into this intersection, opening avenues for further investigation and real-world application. The integration of Generative AI and data quality management stands to offer long-term strategic benefits for organizations, affirming the premise that the future of data quality may well be shaped by advancements in artificial intelligence.

ACKNOWLEDGMENT

We express our heartfelt gratitude to the individuals who have been pivotal in shaping our journey throughout the process of crafting this research paper. Their unwavering support and guidance have been instrumental in bringing this endeavor to fruition. We extend our sincerest thanks to research community who have placed more related content at the internet. Without those content it was impossible to complete research on this topic.

We want to thank our Data Quality Team and Five Below Senior leadership for acknowledging Data Quality value to our organization.

Special mention is reserved for my wife, Beena Dhoni, whose unwavering support, time, and encouragement have empowered me to pursue my passions and excel in this academic pursuit. Lastly, we extend our sincere appreciation to the creators of ChatGPT, whose assistance in refining grammar and sentence structure has been invaluable in enhancing the quality of our paper.

REFERENCES


Pan Singh Dhoni: 1211 Denton Drive, Chester Spring, PA, 19425, USA
Masters in science (M. Sc) Mathematics 2000, Kumaun University, Nainital, Uttrakhand, India
Master Diploma (M. Tech) in Computer Science 2008, National Institute of Electronics, and Information Technology (NIELIT), Delhi, India
Technical Manager, Data Platform, Data Science & Analytics at Five Below Inc. Close to two-decade experience in IT and computer related field.
Email: ps.dhoni@gmail.com