Faculty of Business Economics

Master of Management

Master's thesis

Evaluation of modern tools for data scientists

Oussama Achraf

Thesis presented in fulfillment of the requirements for the degree of Master of Management, specialization Data Science

SUPERVISOR :
Prof. dr. Benoit DEPAIRE

2024
2025

UHASSELT
KNOWLEDGE IN ACTION
www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

Faculty of Business Economics

Master of Management

Master's thesis

Evaluation of modern tools for data scientists

Oussama Achraf

Thesis presented in fulfillment of the requirements for the degree of Master of Management, specialization Data Science

SUPERVISOR :

Prof. dr. Benoit DEPAIRE

Abstract

Data cleaning plays a pivotal role in the data science pipeline, underpinning the reliability and accuracy of data-driven insights. As data volumes grow exponentially, traditional manual cleaning methods have become increasingly inefficient, underscoring the need for advanced automated solutions. Recent advancements in artificial intelligence, particularly large language models (LLMs) like ChatGPT, introduce new possibilities for automating data-cleaning processes. This thesis evaluates the effectiveness of ChatGPT-4 and its variant, ChatGPT-4o, in automating data-cleaning tasks, focusing on key metrics of accuracy, precision, and recall. Through 12 tests across datasets with varying error rates (10% and 30%) and different prompt variants, this study assesses each model's ability to identify and correct data inconsistencies, particularly in textual data, where LLMs are known to excel. Our findings reveal that while ChatGPT models can handle complex data-cleaning tasks and significantly improve efficiency, they still require human oversight to ensure precision, especially in specialized domains. This thesis contributes to the understanding of LLMs 'applicability in data cleaning, offering insights into their strengths, limitations, and potential integration within data science workflows.

1. Introduction

Data cleaning is a critical step in the data science workflow, essential for ensuring the integrity and reliability of data-driven insights [5]. With the exponential growth of data generated from diverse sources and in varying formats, traditional manual data cleaning methods have become increasingly time-consuming and cumbersome. Consequently, there is a growing need for advanced tools to automate various aspects of Extract, Transform, and Load (ETL), data cleaning included, thereby improving efficiency, scalability, and reducing processing time [2].

Recent advancements in artificial intelligence, particularly in large language models (LLMs) like ChatGPT, have opened new avenues for automating tasks [15]. These AI-based tools offer the potential to handle complex data cleaning operations, reduce the burden on data scientists, and enhance overall workflow productivity [4].

This thesis evaluates the effectiveness of LLM-based tools, specifically ChatGPT versions 4 and 4o, in assisting data scientists with automating data cleaning tasks. The evaluation focuses on three key metrics: accuracy, precision, and recall, aiming to determine ChatGPT's viability as a tool for enhancing data cleaning processes in data science workflows.

The main contributions of this thesis include a detailed assessment of ChatGPT's capabilities as a large language model (LLM) for automating data cleaning, along with practical recommendations for its integration into data science workflows. These findings will contribute to the growing knowledge of AI applications in data science, offering valuable insights for researchers and practitioners.

The structure of this thesis is as follows: The *Theoretical Background* section provides a narrative review of the existing literature on data science, data cleaning, and the evolution of data cleaning tools. The *Methodology* section describes the study's experimental design to evaluate ChatGPT-4 and ChatGPT 4-Omni as tools for data cleaning. The *Results* section reports the study's findings, followed by the *Discussion* section, which addresses the implications of these findings. Finally, the thesis concludes with a *Conclusion* section.

2. Theoretical Background

2.1. Foundations of Data Science and the Role of Data Cleaning

Data Science and Its Workflow

Data science is an interdisciplinary field that has become essential for harnessing the vast and complex amounts of data generated in our increasingly digital society [33]. It combines principles from statistics, data mining, databases, and distributed systems to unlock insights and value from data, tackling challenges posed not only by data volume (often termed "Big Data") but also by the complexity and variety of questions data can address [43]. The shift from an "analogue" to a "digital" society has transformed business, communication, and decision-making. Characterised by the "Internet of Things," this transformation involves continuous data collection, leading to the "Big Data" phenomenon [33]. Through a blend of statistical techniques, machine learning algorithms, and domain expertise, data science empowers organisations to make informed decisions, optimise processes, and explore new opportunities, such as in the supply chain design and management domain [46]. This interdisciplinary nature of data science underscores not only the importance of technical skills but also the necessity for clear communication and teamwork to translate complex findings into actionable strategies [48].

A widely recognized framework in data science projects is the CRISP-DM (Cross-Industry Standard Process for Data Mining), which includes business understanding, data understanding, data preparation, modelling, evaluation, and deployment [39]. Exploratory Data Analysis (EDA), the process of exploring patterns, distributions, relationships, and anomalies within the data, is a critical step in data understanding [22]. Data cleaning, a crucial step within the data preparation phase, involves identifying and correcting errors, inconsistencies, and inaccuracies to ensure the quality and reliability of the data. Modelling involves building predictive or descriptive models based on the analyzed data. Finally, deployment integrates these models into a production environment for decision-making processes or operational systems [47].

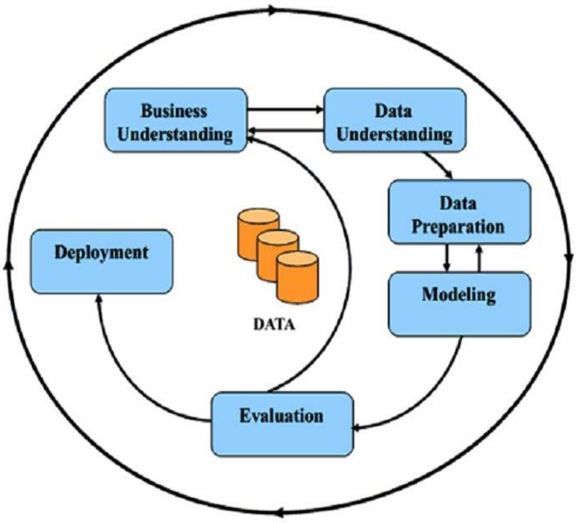


Fig 1: CRIPS-DM framework

Data Cleaning as Part of Data Preparation in Data Science

Challenges in Data Cleaning

Data cleaning is a fundamental step in the data science workflow, essential for ensuring the accuracy and reliability of data-driven insights. It involves identifying and correcting errors, handling missing values, resolving inconsistencies, and managing outliers in the data [5]. Effective data cleaning establishes a foundation of accurate, complete, and consistent data for all subsequent analytical activities. Without thorough data cleaning, data quality may be compromised, leading to incorrect analyses and misleading results — illustrating the principle of 'garbage in, garbage out. [5]. High-quality data is crucial for effective data analysis and modelling, as inaccurate or incomplete data can distort findings, resulting in poor decision-making and potentially significant financial losses for businesses [34]. For instance, predictive models trained on poor-quality data—e.g. resulting from inadequate cleaning techniques—can produce misleading insights, undermining the reliability of business intelligence systems and diminishing the actionable value of their outputs.

While data cleaning is critical for enhancing the performance and accuracy of data analysis and modelling efforts, it is also time-consuming and labour-intensive, often accounting for a large portion of the time spent on data analysis projects [5].

Data cleaning is a crucial step in the data science workflow, yet it involves numerous challenges that complicate the process and can require integrating data from multiple sources and ensuring data privacy and security [12]. While all these challenges are uniquely important, this study will focus on the primary challenge of resolving inconsistencies in textual data, which is arguably one of the most complex aspects of data cleaning and can significantly impact the quality and reliability of data-driven insights.

Textual data often contains a wide range of inconsistencies. These can include extra whitespace (leading or trailing), typos, misspellings, inconsistent use of abbreviations and acronyms, etcetera. Formatting differences, such as variations in date formats (e.g. "11−12−2024" vs "12/11/24") and phone numbers (e.g. (123) 456-7890 vs 123-456-7890), and case-sensitive discrepancies (e.g., "New York" vs "new york"), are also common. Additional issues may involve inconsistent punctuation, encoding differences (like UTF-8 vs. ASCII), variations in metric units (e.g., "kg" vs. "kilograms" or "ml" vs. "millilitres"), and even emojis. Brand names or common phrases might appear in varying ways, such as "CocaCola" vs. "Coca-Cola."

Catching all these errors can be difficult, and anticipating every type of inconsistency can be challenging. Cleaning them often requires a programmatic approach, which can be tedious as it involves inspecting the dataset to identify potential errors and the regex patterns to use to systematically address many of these issues. Moreover, addressing these challenges often requires a combination of technical solutions and domain expertise, including a deep understanding of the data's schema to ensure that inconsistencies are resolved accurately, comprehensively, and in a manner that prepares the data for effective analysis [12].

Data Cleaning Techniques

Data cleaning techniques range from basic manual approaches to advanced automated processes, each suited to different types and scales of data issues. First, data should be loaded and profiled to analyze its distributions, identify error patterns, determine their frequency, and pinpoint the specific columns affected within structured datasets. [6, 30]. For very large datasets, batch processing or distributed computing is often necessary. This might involve using the ***chunksize*** parameter in ***pandas.read_csv()*** in Python to process data in smaller, manageable portions or employing Apache Spark or PySpark for distributed computing to handle the workload of large datasets [6].

Textual errors and inconsistencies, which often follow specific patterns, can be addressed with regex-based cleaning. Since this approach may involve creating regex patterns to address various types of errors, it can become cumbersome and time-consuming for large datasets especially those with a high variance of data error types. Therefore, automating the cleaning process through scripted pipelines, such as those in ETL workflows, is particularly beneficial [47].

After automated cleaning, the data can be manually inspected, or it can undergo automated validation using regular expressions and mapping dictionaries to check for successful cleaning and validate the process's effectiveness and veracity. In many industries, it's crucial to document each step of the cleaning process for reproducibility and transparency [47]. This can be done by logging each action in Python using logging libraries, ensuring a clear and accessible record of the entire data-cleaning process.

Established tools such as Airflow, Luigi, and Prefect are ideal for creating automated data-cleaning workflows since they are designed to orchestrate Extract, Transform, and Load (ETL) pipelines. While they do not clean data directly, they excel in managing and automating the sequence of actions in a pipeline, such as running data extraction, transformation scripts, or integrating with other cleaning tools [50]. However, OpenRefine and Trifacta include built-in data cleaning support and are specifically designed to handle data transformation chores. These tools simplify the process with features like clustering and transformation suggestions, which reduce the need to write new regex scripts. Trifacta also uses machine learning to identify cleaning actions based on data patterns, thus both technologies are excellent options for easing data preparation without requiring considerable manual coding [50].

2.2. Evolution of Data Cleaning Tools

Traditional Tools

Before the rise of advanced AI-driven tools, traditional data-cleaning methods were the standard way of preparing data for analysis. Some of the most widely used traditional tools include Excel and Structured Query Language (SQL) [27].

Microsoft Excel became a go-to choice for data management and cleaning due to its easy accessibility and user-friendly interface. Excel's functions, such as filtering, conditional formatting, like pivot tables, and data validation, make it especially helpful for identifying and correcting errors in smaller to moderately sized datasets.

Python libraries, especially Pandas, brought significant power and performance to data manipulation and cleaning. Pandas library provides a range of functions for handling missing values, merging datasets, and applying complex transformations quickly [6]. Its versatility and seamless integration with other tools such as application program interfaces (APIs) make it a favourite among data scientists for handling various data-cleaning tasks.

Structured Query Language (SQL) is essential for managing and querying data in relational databases. SQL enables users to perform data extraction, aggregation, and updates using commands like SELECT, UPDATE, and DELETE. Its robust querying capabilities allow users to efficiently manage large datasets and handle data cleaning operations directly within databases. With textual data, NoSQL databases (such as MongoDB and Cassandra) are often the go-to choice because they support semi-structured and unstructured data formats. MongoDB, for example, utilises a document-oriented model, making it best for working with textual data with varied fields and formats. Also, these NoSQL databases natively support a distributed architecture and horizontal scalability, which are crucial for managing large datasets efficiently [14].

Limitations of Traditional Tools

Traditional data cleaning tools like Excel and SQL have historically served as mainstays in data preparation, but they come with notable limitations [27]:

- **Scalability Challenges**: These tools often struggle with large-scale datasets. Excel, for example, can quickly become slow and unresponsive beyond a certain data size. While Pandas can handle larger datasets than Excel, it still faces performance issues with very large datasets due to its in-memory processing structure.
- **High Manual Effort**: Using traditional tools for tasks like finding duplicates, resolving inconsistencies, and handling missing data often involves repetitive, manual steps. This not only makes data cleaning labour-intensive but also leaves room for human error [27].
- **Complexity and Learning Curve**: While SQL is powerful for data management, it requires a high level of familiarity with databases, making it difficult for users without experience in database systems. Similarly, Python libraries like Pandas require coding knowledge, which can be a barrier for those without programming skills.
- **Limited Advanced Features**: Traditional tools often lack features like automated anomaly detection, sophisticated imputation methods, and machine learning integration. This means users must rely on additional manual work or multiple tools to achieve a complete data-cleaning process.

As effective as these tools have been, their limitations underscore the need for more advanced, scalable, and automated solutions—especially in the era of big data. These gaps are driving the rise of modern AI-powered data cleaning tools, which bring enhanced capabilities for handling large and complex datasets.

Modern AI-Driven Solutions in Data Cleaning

The rise of Artificial Intelligence (AI) and Machine Learning (ML) has brought transformative advancements to data cleaning. These technologies introduce sophisticated automation capabilities that reduce the manual effort typically required and enhance the speed and precision of data-cleaning tasks. AI and ML algorithms can automatically identify and fix errors, impute missing values, detect duplicates, and resolve inconsistencies in large datasets [8]. Through pattern recognition and predictive modelling, they tackle complex data issues that traditional methods often struggle to handle effectively [8]. For example, machine learning models trained on historical data can predict and accurately fill in missing values, often outperforming simple techniques like mean or median substitution [8].

One major advantage of using AI and ML for data cleaning is their adaptability and continuous improvement. As these models process more data and encounter various data issues, they learn to refine their accuracy, enabling them to handle even more intricate data-cleaning tasks over time [16, 20]. This adaptability makes AI-driven approaches far more dynamic and effective than static, rule-based systems.

Additionally, AI and ML integrate seamlessly with data processing pipelines, supporting automated data cleaning within broader data workflows [47]. This integration ensures high-quality, reliable data for downstream analysis and modelling, ultimately improving overall data quality. Modern AI-powered data cleaning tools, such as Trifacta and IBM Watson, exemplify these capabilities by providing comprehensive platforms for automating data preparation and transformation.

Large Language Models (LLMs) for Data Cleaning

Large Language Models (LLMs) are advanced machine learning models built on Natural Language Processing (NLP) principles, designed to process (or "understand") and generate human-like text. These models use a transformer architecture, trained on massively large datasets that require significant computational resources. This extensive training helps LLMs capture language patterns, including syntax, grammar, facts, and some reasoning capabilities [44]. However, their apparent "reasoning" isn't due to true understanding; instead, it results from statistical predictions. LLMs calculate the most probable next word or sequence based on patterns in language and probability, not from actual comprehension [49].

Because LLMs learn from vast amounts of real-world data, they can also inherit biases present in this data, making their responses and performance highly dependent on the quality and diversity of their training datasets. Notable examples of LLMs include GPT (Generative Pre-trained Transformer) by OpenAI, BERT (Bidirectional Encoder Representations from Transformers) by Google, and T5 (Text-To-Text Transfer Transformer), also by Google [49]. This study will focus on ChatGPT, the latest and most widely used model in the GPT series, making it one of the most popular large language models (LLMs) available today.

ChatGPT

By leveraging the NLP capabilities of Large Language Models (LLMs) and their pattern recognition (learning complex relationships and structures in text), ChatGPT can automate many data-cleaning tasks that traditionally require manual intervention. ChatGPT can assist in automating error detection and correction, standardising data formats, providing intelligent suggestions for data imputation, and enhancing overall data quality.

However, while ChatGPT shows promise in data cleaning, certain challenges and limitations must be considered. As an NLP-based model trained on a vast but generic dataset—predominantly textual data—it was not specifically designed for data-cleaning tasks. While it can suggest code for cleaning operations when prompted with a dataset, its general-purpose training may limit its effectiveness and consistency for specialized data-cleaning needs [36, 44].

Additional training of transformers with vast amounts of domain-specific data would be required for specific-domain tasks, including specialised data-cleaning needs [49]. This demands significant computational resources, which could pose challenges for deployment in resource-limited environments. Although ChatGPT can automate many tasks, human oversight is still necessary to ensure the correctness and relevance of cleaned data [36, 44]. These challenges underscore the need for empirical evaluation to fully understand the capabilities and limitations of ChatGPT in the context of data cleaning.

Impact

To assess the effectiveness of modern data cleaning tools, it is essential to establish clear evaluation criteria, with key metrics being **performance**, **accuracy**, and **user-friendliness**.

- **Performance (Speed and Efficiency)**: AI-based tools excel at reducing repetitive manual intervention, significantly enhancing the speed of data cleaning tasks, especially for large datasets. They are particularly faster at cleaning unstructured textual data compared to traditional tools, thanks to their NLP capabilities specifically designed to handle and process text efficiently [44]. By automating time-intensive processes, these tools free up resources for critical analytical and interpretative work, allowing data projects to progress more rapidly. This efficiency enables faster turnaround times, providing timely insights and supporting quicker decision-making.

  However, Large Language Models (LLMs), including ChatGPT, can encounter scalability challenges with large datasets, similar to traditional tools. LLMs have a context size limitation, determined by the number of tokens they can process at once. For instance, the standard ChatGPT-4 model handles up to 8,192 tokens (about 6,000 words) [49]. This constraint means that LLMs may not necessarily offer a performance advantage over traditional tools when processing extensive datasets, as they must work within these token limits.
- **Accuracy:** While LLMs can be efficient for cleaning unstructured textual data, their precision in the data-cleaning process may not surpass that of traditional tools. Traditional tools are specifically designed and fine-tuned for data cleaning tasks, often making them more precise, especially for structured, domain-specific data. LLMs are trained on broad, general-purpose datasets, so they may lack the specialized knowledge that data scientists need when working with data from a particular domain. Fine-tuning allows LLMs to adapt from general-purpose models to more specialized tasks [36].

  The training of LLMs with vast amounts of general data makes them, nonetheless, highly adaptable to varied fields, which is a strength that compensates for their lack of domain-specific precision. LLMs can learn patterns within their context window, improving their effectiveness with consistent usage.
- **User-Friendliness (Ease of Use and Overall User Experience):** Most LLMs are designed with intuitive prompting systems, often providing suggestions and guidance based on the user's input, making the interaction feel almost like conversing with another person. This approach greatly enhances the usability of data-cleaning tools by reducing cognitive load and guiding users through complex tasks that traditionally require programming expertise [16]. A user-friendly tool like this lowers the learning curve, enabling aspiring data scientists and analysts to perform cleaning tasks efficiently without needing extensive training.

  Modern AI-driven tools are designed to integrate smoothly with existing data processing pipelines and platforms, creating a cohesive workflow from data collection through to analysis. This seamless integration can enhance data management practices across the board. By incorporating automated data cleaning into the broader data workflow, organizations can ensure that high-quality data is readily available for all downstream processes, improving consistency and reliability throughout.

3. Methodology

Data cleaning, particularly with unstructured data, often consumes a significant portion of a data scientist's time, detracting from the more critical tasks of analysis and insight generation [40]. This study adopts a quantitative approach to evaluate ChatGPT's effectiveness in streamlining data cleaning, to reduce time spent on this labour-intensive process and enable professionals to focus on producing actionable insights.

The study will compare the effectiveness of ChatGPT-4 and its variant (ChatGPT-4-Omni, or 4o) in data cleaning, using accuracy as the primary metric and assessing precision and recall. The main distinction between ChatGPT-4 and ChatGPT-4-Omni lies in their design: ChatGPT-4 is built for highly demanding tasks that require high quality and precision in its responses, excelling in capturing nuanced details. In contrast, ChatGPT-4-Omni is a lighter, more computationally efficient version designed for everyday tasks. While it is faster and less resource-intensive, it may offer slightly lower precision than ChatGPT-4.

3.1 Study Design

The research design uses a clean dataset sourced from Kaggle, specifically the **‘Top Box-Office Movies - Analysis’** dataset, titled ***boxofficemojotopfranchises.tsv*** (15.25 kB).

Using Python, the dataset was cloned twice, and random errors were introduced into the cloned datasets at rates of **10%**and **30%,** respectively, targeting the columns containing string variables. The percentage of errors was calculated based on the number of indices in which errors were introduced, rather than the total number of errors within each column.

The errors introduced fell into the following categories:

1. **Typographical errors**: This included misspellings, missing characters, and transposed characters.
2. **Inconsistent formatting**: Issues such as case sensitivity, inconsistent punctuation use, and trailing/leading whitespace were introduced.
3. **Abbreviations and Expansions**: Synonyms, inconsistent abbreviations, and expansions were also introduced to create variation.

The modified datasets were then uploaded to both **ChatGPT-4** and **ChatGPT-4-Omni** (referred to as ChatGPT-4o). The following prompts were used for both models to ensure consistency:

- "Identify all indices where the string data might need cleaning."
- "Locate all indices with inconsistent or erroneous string data in this dataset."
- "Conduct a thorough analysis of string inconsistencies and errors in the string columns, and identify all corresponding indices.”

Importantly, after each prompt, the ChatGPT history is cleared to prevent the retention of previous interactions that could influence the results of subsequent prompts.

Since ChatGPT is a large language model (LLM) based on natural language processing (NLP), it cannot directly execute code or run a local programming environment to clean data automatically [23]. Instead, it generates code that must be executed externally on the dataset.

The scripts were executed in VS Code, and upon running the script, the results − which identify indices with potential errors − were compared to the ground truth (i.e., the dataset where errors were intentionally imputed for testing).

Data

The **Top Box-Office Movies - Analysis** dataset (***boxofficemojotopfranchises.tsv***) was chosen for several key reasons:

1. **Real-World Relevance**: As a widely accessible, real-world dataset, it provides a more authentic testing environment compared to simulated datasets.
2. **Manageable Size**: With fewer than 1,000 rows, the dataset is small enough to be easily processed and uploaded into ChatGPT, while still containing a diverse mix of both string and numerical columns.
3. **Simple**: The dataset consists of four columns, containing both string-based and numerical data. It serves as an ideal simple dataset for introducing a variety of string inconsistencies, including typographical errors, inconsistent formatting, and abbreviations.

3.2 Data Analysis

The results from the research study were generated from two distinct models: **ChatGPT-4** and **ChatGPT-4o**. Each model was evaluated using two experimental conditions based on datasets that had been intentionally error-imputed at rates of **10%** and **30%**. Furthermore, for each model and error rate combination, three different prompts were utilised, providing replicates to ensure the robustness of the findings, therefore giving a total of 12 tests in total.

Model	Error Rate	Prompt 1	Prompt 2	Prompt 3
ChatGPT−4	10%	✓	✓	✓
	30%	✓	✓	✓
ChatGPT−4o	10%	✓	✓	✓
	30%	✓	✓	✓

The primary evaluation metric for assessing the correctness of the data cleaning output across the groups was **accuracy**, resulting in a total of 12 accuracy scores. Additionally, **precision** and **recall** were reported to provide a more comprehensive analysis of performance.

$$\textbf{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

$$\textbf{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

$$\textbf{Accuracy} = \frac{\text{True Positive + True Negative}}{\text{Total}}$$

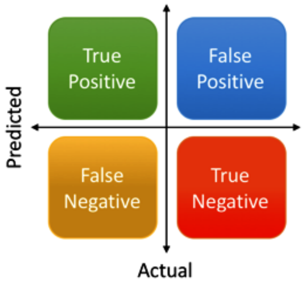


Fig 2: Metrics

Given the small sample size, the presence of more than three independent groups, and the need for a non-parametric test, the **Kruskal-Wallis H Test**, an alternative to the parametric ANOVA, was employed for statistical analysis.

To address the issue of multiple comparisons, the **Benjamini-Hochberg (BH) method** was applied to control the **false discovery rate (FDR)**. This method is less conservative than stricter corrections like the Bonferroni correction, making it more suitable for exploratory research, where maintaining the power to detect true effects is essential.

Finally, **post-hoc testing** was conducted to determine significant differences among the groups, particularly between the models (ChatGPT-4 and ChatGPT-4o) and the varying error rates.

4. Results

Upon uploading the dataset and providing the prompt, the GPT models consistently identified the 'Movie 'column as the sole string column. Notably, the models returned messages indicating that they had detected string inconsistencies. While the types of inconsistencies varied, leading and trailing whitespaces were the most frequently encountered issues. A follow-up prompt was necessary to generate a downloadable CSV file containing the identified problematic rows for further analysis.

The flagged rows were compared against the rows with deliberately introduced string inconsistencies (i.e., the ground truth). Based on this comparison, a table was generated to categorize the results into true positives, false positives, true negatives, and false negatives. The resulting confusion matrices are shown below:

GPT-4o _ 10% Error Rate

		Predicted - Prompt 1		**Predicted - Prompt 2**		**Predicted - Prompt 3**	
		1	**0**	**1**	**0**	**1**	**0**
Actual	**1**	38	62	15	85	78	22
	0	219	681	29	871	297	603
		1 =	Positive				
		0 =	Negative				

GPT-4o _ 30% Error Rate

		Predicted - Prompt 1		**Predicted - Prompt 2**		**Predicted - Prompt 3**	
		1	**0**	**1**	**0**	**1**	**0**
Actual	**1**	245	55	45	255	82	218
	0	568	132	46	654	59	641
		1 =	Positive				
		0 =	Negative				

GPT-4 _ 10% Error Rate

		Predicted - Prompt 1		Predicted - Prompt 2		Predicted - Prompt 3	
		1	0	1	0	1	0
Actual	1	33	67	44	56	33	67
	0	45	855	219	681	153	747
		1 =	Positive				
		0 =	Negative				

GPT-4 _ 30% Error Rate

		Predicted - Prompt 1		Predicted - Prompt 2		Predicted - Prompt 3	
		1	0	1	0	1	0
Actual	1	100	200	121	179	224	76
	0	109	591	92	608	247	453
		1 =	Positive				
		0 =	Negative				

Based on the confusion matrices above, the key metric of focus—accuracy in ChatGPT's ability to identify the deliberately introduced string inconsistencies—was evaluated. Additionally, precision and recall were calculated to provide a more comprehensive performance analysis. The tables below present these metrics:

Accuracy scores

Model	Error Rate	Prompt 1	Prompt 2	Prompt 3
ChatGPT−4	10%	0.888	0.725	0.78
	30%	0.691	0.729	0.677
ChatGPT−4o	10%	0.719	0.886	0.681
	30%	0.377	0.699	0.723

Precision scores

Model	Error Rate	Prompt 1	Prompt 2	Prompt 3
ChatGPT−4	10%	0.423	0.167	0.177
	30%	0.478	0.568	0.476
ChatGPT−4o	10%	0.148	0.341	0.208
	30%	0.301	0.495	0.582

Recall scores

Model	Error Rate	Prompt 1	Prompt 2	Prompt 3
ChatGPT−4	10%	0.330	0.440	0.330
	30%	0.333	0.403	0.747
ChatGPT−4o	10%	0.380	0.150	0.780
	30%	0.817	0.150	0.273

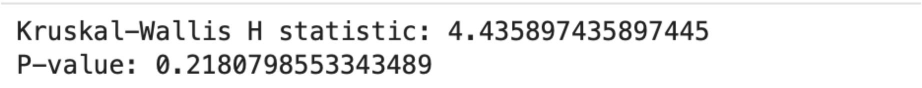
Based on these statistics, the overall average accuracy across the four groups was 71.5% (0.715). Notably, the average accuracies for the 10% error rate datasets in both ChatGPT models were above this overall average, at 79.8% (0.798)[ChatGPT-4] and 76.2% (0.762)[ChatGPT-4o]. In contrast, the 30% error rate datasets showed lower-than-average accuracies, at 69.9% (0.699)[ChatGPT-4] and 60.0% (0.600)[ChatGPT-4o].

In contrast, the datasets with a 30% error rate performed better than average (36.4% or 0.364) in terms of precision compared to those with a 10% error rate. The precision scores for the 30% error rate datasets were 45.9% (0.459)[ChatGPT-4o] and 50.7% (0.507)[ChatGPT-4], whereas the 10% error rate datasets had lower precision scores of 23.2% (0.232)[ChatGPT-4o] and 25.6% (0.256)[ChatGPT-4].

The recall scores across the four groups were relatively consistent, averaging around 42.8%.

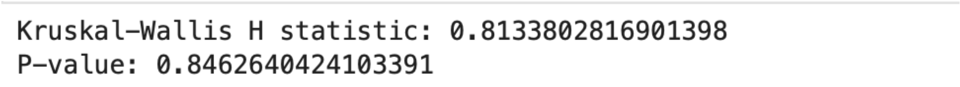
The individual metrics were then grouped based on the ChatGPT model used and the error rate of the inconsistencies introduced into the dataset, resulting in four distinct groups. Each group contained three values corresponding to the three different prompts. To evaluate whether there were significant differences in the

ChatGPT models ʼabilities to identify these inconsistencies, the Kruskal-Wallis H Test was applied. The *stats* module from the *SciPy* library in Python was used to run this test, and the results were as follows;



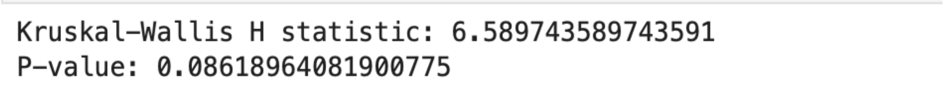
```
Kruskal-Wallis H statistic: 4.435897435897445
P-value: 0.2180798553343489
```

Fig 3: Accuracy



```
Kruskal-Wallis H statistic: 0.8133802816901398
P-value: 0.8462640424103391
```

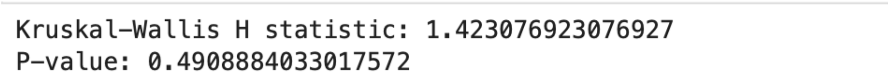
Fig 4: Recall



```
Kruskal-Wallis H statistic: 6.589743589743591
P-value: 0.08618964081900775
```

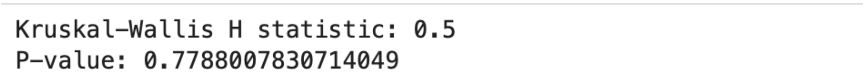
Fig 5: Precision

Although there were fluctuations in the metrics across the prompts, no significant differences were found between any of the four groups in terms of accuracy, precision, or recall scores, using an alpha value of 0.05. This indicates that neither the GPT models used nor the error rates introduced into the dataset (10% or 30%) had a significant impact on the models ʼperformance in these metrics. Consequently, further analysis was conducted to determine whether significant differences existed between the prompts themselves. The results, as shown below, indicated no significant difference (α = 0.05).



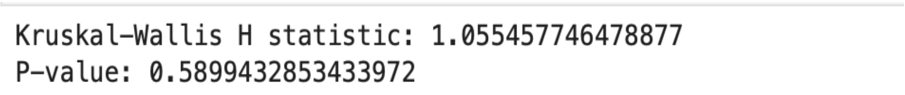
```
Kruskal-Wallis H statistic: 1.423076923076927
P-value: 0.4908884033017572
```

Fig 6: Accuracy



```
Kruskal-Wallis H statistic: 0.5
P-value: 0.7788007830714049
```

Fig 7: Precision



```
Kruskal-Wallis H statistic: 1.055457746478877
P-value: 0.5899432853433972
```

Fig 8: Recall

Given these results, there was no need to conduct a post-hoc analysis for pairwise comparisons, such as using Dunn’s test, which would have required applying multiple testing corrections.

5. Discussion

The results indicate no significant difference between ChatGPT-4 and ChatGPT-4-Omni(4o) in identifying string(textual) inconsistencies for data cleaning tasks. ChatGPT-4, a legacy model from OpenAI, is touted for its superior language and reasoning capabilities compared to earlier versions like GPT-3 and 3.5 [32]. It excels at handling intricate and nuanced prompts, making it ideal for complex tasks. ChatGPT-4-Omni, on the other hand, is a lighter version designed to use fewer resources, offering faster performance while maintaining much of the capabilities of ChatGPT-4 [32]. However, as OpenAI notes, it is not as finely tuned as the larger legacy model.

OpenAI advises that ChatGPT-4 is better suited for detailed, in-depth research tasks, while ChatGPT-4-Omni is more appropriate for everyday tasks, as it prioritizes speed and cost-efficiency over precision and depth [32]. The fact that ChatGPT-4 did not outperform ChatGPT-4-Omni based on statistical analysis suggests that the lighter version may be more advantageous for data cleaning tasks, given its reduced resource consumption and lower computational costs.

The ChatGPT 4 and 4o models achieved around 70% accuracy in identifying string(textual) inconsistencies in the dataset. However, this metric may not be as compelling as it seems. The dataset had a class imbalance, with Class 1 (representing deliberately introduced inconsistencies) making up either 10% or 30% of the data, depending on the introduced error rate. With a 10% error rate, a model that simply predicts Class 0 for the entire dataset (i.e., assuming no inconsistencies) would still be 90% accurate. Similarly, with a 30% error rate, such a model would achieve 70% accuracy. This highlights the need for further more nuanced evaluation, as accuracy alone may not fully capture model performance in imbalanced scenarios.

As a result, despite the seemingly decent accuracy score, the precision and recall scores were underwhelming. The average precision across both models was 36.4%, which is below average, indicating that most of the positive predictions (identifying text with inconsistencies) were inaccurate. Precision was even lower with the 10% error rate dataset due to the more pronounced class imbalance. The average recall score, while slightly higher at 42.8%, was still below average. This means the models had a relatively high rate of false negatives, failing to capture a significant portion of the actual positives. Using the F1 score, which provides a more balanced measure by combining precision and recall, the metric is 39.34%, which is relatively low, falling below 50%.

The results indicate that while ChatGPT performs well with textual data compared to numerical data—owing to its foundation in natural language processing—it remains imprecise in identifying string (textual) inconsistencies, making it unreliable for comprehensive data cleaning tasks. A major finding is the inconsistency in performance metrics across different prompts. For instance, in one test (ChatGPT-4o with a 30% error rate), there was a 34% variation in accuracy between the lowest and highest prompt accuracy scores. Similar discrepancies were observed in precision and recall, with one test showing a substantial 63% gap in recall, highlighting the model's variability in handling these tasks.

The non-deterministic nature of ChatGPT's results stems from its use of probabilistic methods to generate outputs. This approach can pose challenges for automating data cleaning tasks, where consistency and reproducibility are critical. Automated data cleaning, especially within workflow pipelines, requires reliable, repeatable outcomes—something traditional tools, powered by programmatically written scripts, excel at. ChatGPT’s outputs are more conversational, relying on probabilistic predictions to ‘guess ’the next step based on previous inputs and context. This lack of precision and consistency makes it less suitable for data cleaning, which demands accuracy and uniformity.

More so, this inconsistency in generating results highlights the need for quality assurance to thoroughly check and validate ChatGPT outputs [37]. Without this step, there is a higher risk of errors and misinterpretations, which could negatively impact the downstream data-cleaning process. In turn, these issues may cascade into downstream data analysis or modelling, leading to unreliable or inaccurate outcomes.

Another issue is ChatGPT's token limit, which restricts the amount of text it can process in a single interaction while still maintaining context [32, 49]. This limitation means it cannot effectively handle large datasets or retain a broader context when repeatedly prompted with extensive data. As a result, this study chose a manageable dataset for ChatGPT, limiting the size to under 1,000 rows with only a few columns. This issue also highlights the scalability challenges of using ChatGPT for data cleaning, where traditional tools significantly outperform the current ChatGPT models evaluated in this research.

It is important to keep in mind that ChatGPT models are trained on vast amounts of general, non-domain-specific information. For applications that require high precision in a specific domain, like medicine, it is advisable to fine-tune the model transformers with a specialized dataset. Domain-specific training helps improve accuracy and reduce ambiguity that can arise from general-purpose training [49]. For example, in the medical field, the term "fall" often refers to a patient's physical fall. Without domain-specific training, however, the model might interpret "fall" as the autumn season due to its polysemous nature, potentially leading to misunderstandings (e.g., associating "fall" with the timing of treatment or discharge instead of a patient's accident).

Despite its limitations, ChatGPT remains invaluable in data cleaning, particularly for unstructured datasets with significant textual content, especially when a data scientist is still exploring the raw data. In these exploratory stages, ChatGPT proves beneficial by gathering contextual insights about the dataset, even without explicit information about its contents [16]. For instance, in this case study, without prior knowledge, the ChatGPT models were able to infer that the dataset was related to movies.

Cleaning unstructured and messy data is often challenging, and data scientists may struggle with where to begin. Programmatically, they would need to create complex rules and regular expressions, which might still fall short of covering every possible textual inconsistency. In contrast, ChatGPT models offer flexibility and adaptability for handling such messy data, making them an excellent starting point for data cleaning [16]. This flexibility makes ChatGPT an effective tool for rapid prototyping and data exploration in the early stages, streamlining the process of understanding and organizing raw data.

5.1 Limitations

A major limitation of this study was the sample size. By examining only 10% and 30% error rates in the datasets and using three distinct prompts with ChatGPT, the sample size was relatively small, affecting the robustness of the findings. Additionally, the experiment's outcomes were highly influenced by the quality of the prompts, a dependency that impacts the generalizability of these results. While variability was observed across prompts (though not statistically significant), more prompts might have increased the likelihood of detecting significant differences or trends.

The range of textual inconsistencies introduced into the dataset was also limited, meaning that the study did not capture the full spectrum of inconsistencies typically found in raw data. Expanding the types and variety of inconsistencies would better reflect the complexity of real-world datasets and allow for a more comprehensive assessment of ChatGPT models' capabilities in handling such challenging data.

6. Conclusion

In conclusion, this study demonstrates that while large language models (LLMs) like ChatGPT bring unique advantages to data cleaning, they also come with limitations when compared to traditional approaches. Due to their probabilistic, natural language processing (NLP)-based responses, LLMs can produce variable outputs for similar prompts, leading to inconsistent results. Although our findings showed this variability wasn’t statistically significant, it highlights a challenge for reproducibility—an area where traditional, rule-based methods excel by providing stable, repeatable results. Thus, traditional tools still outperform LLMs for tasks requiring high reproducibility, whose fluctuations could affect downstream data analyses.

However, LLMs like ChatGPT shine in exploratory data cleaning, particularly for unstructured or text-heavy data. Their NLP foundation allows them to identify a wide range of string inconsistencies with flexibility, even without highly specific prompts, making them ideal for the early stages of data exploration. In this study, ChatGPT’s ability to interpret the dataset contextually and pinpoint relevant columns and inconsistencies underscores its strengths in scenarios where adaptability and interpretive power are prioritized over strict reproducibility.

In essence, both LLMs and traditional data cleaning methods offer distinct strengths and trade-offs. Traditional methods ensure reproducibility through explicit, rule-based programming, though they may fall short of catching all text inconsistencies. In contrast, LLMs offer the flexibility to detect a variety of inconsistencies and infer context within the data, facilitating a deeper understanding of the dataset’s structure. Rather than viewing LLMs as replacements for traditional data-cleaning tools, they should be seen as complementary, each fulfilling distinct roles within the data-cleaning process.

Moreover, ChatGPT’s token and context limitations pose challenges for scalability, especially with larger datasets commonly used in data science. Future advancements with expanded context windows and token capacities may better address these needs. This study underscores that a robust hybrid approach—leveraging both AI and traditional methods—is the most effective data-cleaning strategy. AI can speed up the initial stages of data cleaning by quickly capturing context, which can be difficult for rule-based tools. By combining the contextual agility of AI with the precision and consistency of traditional methods, we can create a data-cleaning workflow that is comprehensive, efficient and well-suited to the complexities of modern data science.

Bibliography

1. Abedjan, Z., Chu, X., Deng, D., Fernandez, R. C., Ilyas, I. F., Ouzzani, M., Papotti, P., Stonebraker, M., & Tang, N. (2016). Detecting data errors: Where are we and what needs to be done? *Proceedings of the VLDB Endowment*, *9*(12), 993–1004. https://doi.org/10.14778/2994509.2994518
2. Agarwal, R., & Dhar, V. (2014). Editorial—Big data, data science, and analytics: The opportunity and challenge for IS research. *Information Systems Research, 25*(3), 443–448. https://doi.org/10.1287/isre.2014.0546
3. Aggarwal, C., Bouneffouf, D., Samulowitz, H., Buesser, B., Hoang, T., Khurana, U., Liu, S., Pedapati, T., Ram, P., Rawat, A., Wistuba, M., & Gray, A. (2019). How can AI automate end-to-end data science? *arXiv*.http://arxiv.org/abs/1910.14436
4. Biester, F., Abdelaal, M., & Del Gaudio, D. (2024). LLMClean: Context-Aware Tabular Data Cleaning via LLM-Generated OFDs. *arXiv*. https://doi.org/10.48550/arXiv.2404.18681
5. Chai, C. P. (2020). The importance of data cleaning: Three visualization examples. *CHANCE, 33*(1), 4–9. https://doi.org/10.1080/09332480.2020.1726112
6. Chen, D. Y. (2017). *Pandas for everyone: Python data analysis*. Addison-Wesley Professional.
7. Chu, X., Ilyas, I. F., Krishnan, S., & Wang, J. (2016). Data cleaning: Overview and emerging challenges. In *Proceedings of the 2016 International Conference on Management of Data* (pp. 2201–2206). Association for Computing Machinery. https://doi.org/10.1145/2882903.2912574
8. Côté, P.-O., Nikanjam, A., Ahmed, N., Humeniuk, D., & Khomh, F. (2024). Data cleaning and machine learning: A systematic literature review. *Automated Software Engineering*, *31*(54). https://doi.org/10.1007/s10515-024-00453-w
9. Dasu, T., & Johnson, T. (2003). *Exploratory data mining and data cleaning*. Wiley. https://doi.org/10.1002/0471448354
10. Datrics AI. (2024, June 20). *How to automate data cleaning: A comprehensive guide*. https://www.datrics.ai/articles/how-to-automate-data-cleaning-a-comprehensive-guide
11. Dayal, U., Castellanos, M., Simitsis, A., & Wilkinson, K. (2009). Data integration flows for business intelligence. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*(pp. 1–11). Association for Computing Machinery. https://doi.org/10.1145/1516360.1516362
12. Ganti, V., & Sarma, A. D. (2013). *Data cleaning: A practical perspective*. Morgan & Claypool Publishers.
13. Goth, G. (2015). Bringing big data to the big tent. *Communications of the ACM, 58*(6), 17–19. https://doi.org/10.1145/2771299
14. Gudivada, V. N., Apon, A., & Ding, J. (2017). Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. *International Journal on Advances in Software 10.1* (2017), pp. 1 - 20. https://www.researchgate.net/publication/318432363_Data_Quality_Considerations_for_Big_Data_and_Machine_Learning_Going_Beyond_Data_Cleaning_and_Transformations
15. Haleem, A., Javaid, M., & Singh, R. P. (2022). An era of ChatGPT as a significant futuristic support tool: A study on features, abilities, and challenges. *BenchCouncil Transactions on Benchmarks, Standards and Evaluation, 2*, 100089. https://doi.org/10.1016/j.tbench.2023.100089
16. Hassani, H., & Silva, E. S. (2023). The role of ChatGPT in data science: How AI-assisted conversational interfaces are revolutionizing the field. *Big Data and Cognitive Computing, 7*(2), 62. https://doi.org/10.3390/bdcc7020062
17. Hoyt, R. E., Snider, D. H., Thompson, C. J., & Mantravadi, S. (2016). IBM Watson Analytics: Automating visualization, descriptive, and predictive statistics. *JMIR Public Health and Surveillance, 2*(2), e5810. https://doi.org/10.2196/publichealth.5810
18. IEEE Xplore. (n.d.). *Data cleaning for data quality* | IEEE Conference Publication. Retrieved June 20, 2024, from https://ieeexplore.ieee.org/abstract/document/7724284

19. Kalla, D., Smith, N., & Samaah, F., & Kuraku, S. (2023). Study and analysis of Chat GPT and its impact on different fields of study. *SSRN*. https://papers.ssrn.com/abstract=4402499

20. Kaufmann, T., Weng, P., Bengs, V., & Hüllermeier, E. (2024). A survey of reinforcement learning from human feedback. *arXiv*. https://doi.org/10.48550/arXiv.2312.14925

21. Kanaries. (2023). The ultimate guide to data analysis workflow: Step-by-step. *Kanaries Documentation*. Retrieved June 20, 2024, from https://docs.kanaries.net/articles/data-analysis-workflow

22. Monkman, M. H. (2024). *The data preparation journey*. Chapman and Hall/CRC.

23. Khongrit, A., Limsiri, C., & Meehom, S. (2024). Application of generative artificial intelligence in data cleaning and preparation: A case study of recycled polypropylene composite mixed with tea residue. *Journal of Vongchavalitkul University*, *37*(1). Retrieved from https://ph01.tci-thaijo.org/index.php/vujournal/article/view/257269/172420

24. Khosravi, P., Vergari, A., Choi, Y., Liang, Y., & Van den Broeck, G. (2020). Handling missing data in decision trees: A probabilistic approach. *arXiv*. https://doi.org/10.48550/arXiv.2006.16341

25. Kothuru, S. K., Kumar, V. S., Vadlamudi, A. K., & Rangineni, S. (2023). Analysis on data engineering: Solving data preparation tasks with ChatGPT to finish data preparation. *Journal of Emerging Technologies and Innovative Research, 10*(12), f653–f661. https://www.researchgate.net/publication/377300435_ANALYSIS_ON_DATA_ENGINEERING_SOLVING_DATA_PREPARATION_TASKS_WITH_CHATGPT_TO_FINISH_DATA_PREPARATION

26. Knight, M. (2024). The impact of generative AI on data science. *DATAVERSITY*. https://www.dataversity.net/the-impact-of-generative-ai-on-data-science/

27. Linoff, G. S. (2015). *Data analysis using SQL and Excel*. John Wiley & Sons.

28. Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data*. John Wiley & Sons.

29. Loureiro, A., Torgo, L., & Soares, C. (2004). Outlier detection using clustering methods: A data cleaning application. *ResearchGate*. https://www.researchgate.net/publication/228541549_Outlier_detection_using_clustering_methods_a_data_cleaning_application

30. Mueller, H., & Freytag, J. (2005). *Problems, methods, and challenges in comprehensive data cleansing*. [Conference presentation]. https://api.semanticscholar.org/CorpusID:15756458

31. Oni, S., Chen, Z., Hoban, S., & Jademi, O. (2019). A comparative study of data cleaning tools. *International Journal of Data Warehousing and Mining (IJDWM)*, *15*(4), 48–65. https://doi.org/10.4018/IJDWM.2019100103

32. OpenAI (n.d). "Models." *OpenAI Documentation*. Accessed November 25, 2024. https://platform.openai.com/docs/models.

33. Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big Data*, *1*(1), 51–59. https://doi.org/10.1089/big.2013.1508

34. Provost, F., & Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media, Inc.

35. Rahm, E., & Do, H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin*, *23*(4), 3–13. https://www.researchgate.net/publication/220282831_Data_Cleaning_Problems_and_Current_Approaches

36. Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations, and future scope. *Internet of Things and Cyber-Physical Systems, 3*, 121–154. https://doi.org/10.1016/j.iotcps.2023.04.003

37. Rocha, R. (2024). Using ChatGPT to clean data: An experiment. *Roberto Rocha*. https://robertorocha.info/using-chatgpt-to-clean-data-an-experiment/

38. Ronanki, K., Cabrero-Daniel, B., Horkoff, J., & Berger, C. (2024). Requirements engineering using generative AI: Prompts and prompting patterns. In A. Nguyen-Duc, P. Abrahamsson, & F. Khomh (Eds.), *Generative AI for effective software development* (pp. 109–127). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-55642-5_5.

39. Saltz, J. S. (2021). CRISP-DM for data science: Strengths, weaknesses and potential next steps. In *2021 IEEE International Conference on Big Data (Big Data)* (pp. 2337–2344). IEEE. https://doi.org/10.1109/BigData52589.2021.9671634

40. Singh, S. K., & Dwivedi, D. R. K. (2020). Data mining: Dirty data and data cleaning. *SSRN*. https://doi.org/10.2139/ssrn.3610772

41. Techment (2024, June 20). *Empowering data quality: The role of generative AI in shaping data engineering*. https://www.techment.com/empowering-data-quality-the-role-of-generative-ai-in-shaping-data-engineering/

42. Tu, X., Zou, J., Su, W., & Zhang, L. (2024). What should data science education do with large language models? *Journal Name*. https://doi.org/10.1162/99608f92.bff007ab

43. van der Aalst, W. (2016). Data science in action. In W. van der Aalst (Ed.), Process mining: Data science in action (pp. 3–23). *Springer*. https://doi.org/10.1007/978-3-662-49851-4_1

44. Vangeli, M. (2024). Large language models as advanced data preprocessors: Transforming unstructured text into fine-tuning datasets (UPTEC STS 24032). Uppsala Universitet. Examensarbete, 30 hp. https://uu.diva-portal.org/smash/get/diva2:1879125/FULLTEXT01.pdf

45. vmadhuvarshi (2024). ChatGPT for business process optimization & data cleansing: Blog 3 of series "ChatGPT and SAP." *SAP Community*. https://community.sap.com/t5/technology-blogs-by-members/chatgpt-for-business-process-optimization-data-cleansing-blog-3-of-series/ba-p/13578626

46. Waller, M. A., & Fawcett, S. E. (2013). Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management. *Journal of Business Logistics*, *34*(2), 77–84. https://doi.org/10.1111/jbl.12010

47. Shelf. (2024, September 9). *Data pipelines in artificial intelligence*. Shelf. Retrieved December 14, 2024, from https://shelf.io/blog/data-pipelines-in-artificial-intelligence/

48. Zhang, A. X., Muller, M., & Wang, D. (2020). How do data science workers collaborate? Roles, workflows, and tools. *Proceedings of the ACM on Human-Computer Interaction*, *4*, 1–23. https://doi.org/10.1145/3392826.

49. Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... Wen, J. R. (2023). *A survey of large language models*[ongoing work]. arXiv. https://doi.org/10.48550/arXiv.2303.18223

50. Matskin, M., Tahmasebi, S., Layegh, A., Payberah, A.H., Thomas, A., Nikolov, N., & Roman, D. (2021). A Survey of Big Data Pipeline Orchestration Tools from the Perspective of the DataCloud Project. *International Conference on Data Analytics and Management in Data Intensive Domains*. https://ceur-ws.org/Vol-3036/paper05.pdf