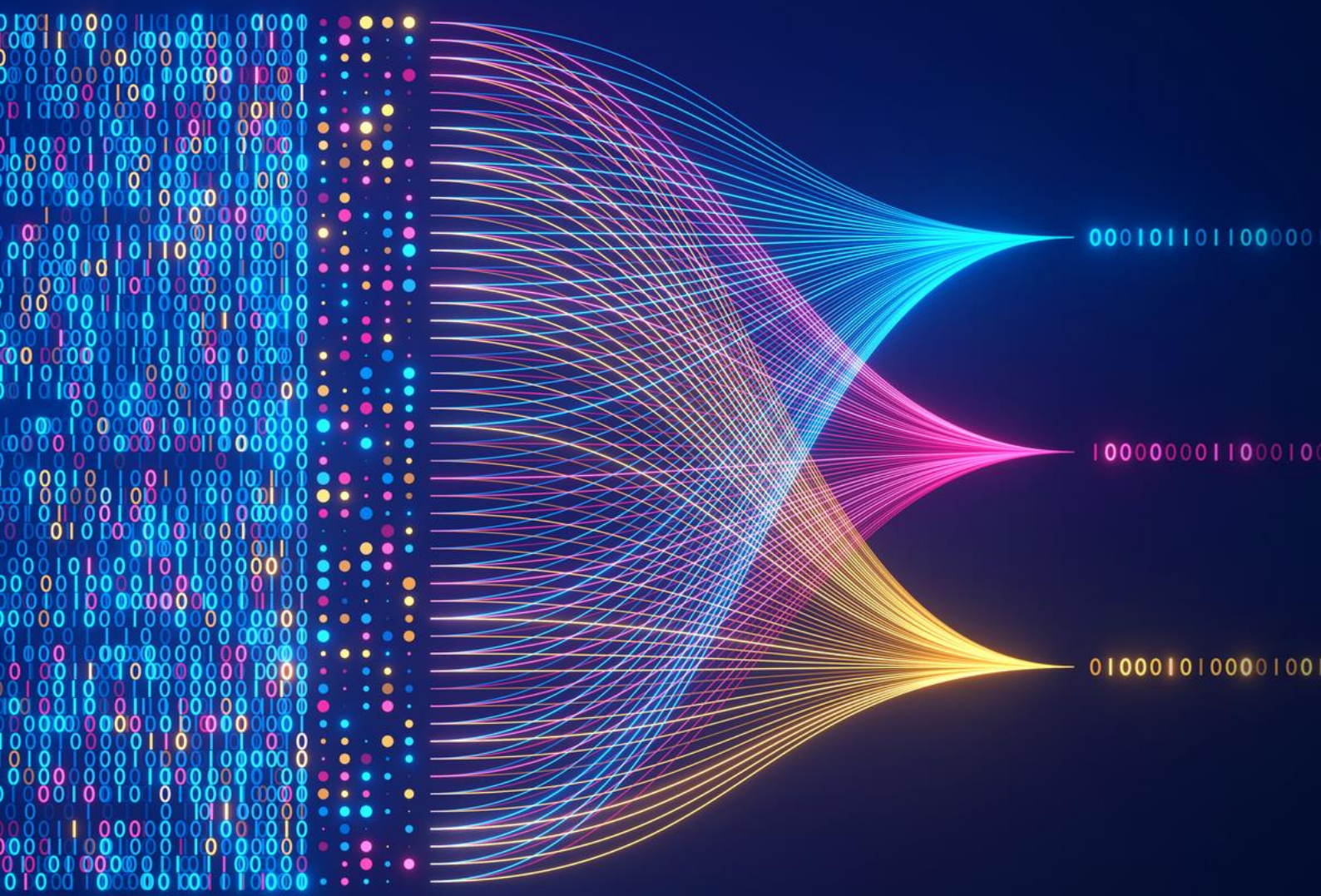
# Assessing the Suitability of AI Tools for Data Correction and Enhancement: An Evaluation and Benchmarking Framework

Master Thesis by M. Stankaitis



TU Delft

*Intended to be empty.*

# Assessing the Suitability of AI Tools for Data Correction and Enhancement: An Evaluation and Benchmarking Framework

Master Thesis

By

M. Stankaitis

**Master of Science**
in Management of Technology

at the Delft University of Technology,
to be defended publicly on 20-02-2025

| | | |
|---|---|---|
| Supervisor: | Dr.ing. V.E. Scholten | |
| Thesis committee: | Dr.ing. V.E. Scholten, | TU Delft |
| | Asst. Prof. Steffen Steinert, | TU Delft |



TU Delft

# Executive summary

This study examines the potential and limitations of generative AI tools, including ChatGPT, Microsoft Copilot, and Google Bard, in data correction and enhancement tasks. While generative AI tools have gained attention for automating data processing, their effectiveness in handling real-world dataset issues such as noise, incompleteness, and inconsistencies remain insufficiently understood. This gap affects users who apply AI for data correction without fully understanding its pitfalls, potentially leading to errors and unreliable outputs. Additionally, researchers lack a common evaluation framework, making it difficult to compare AI performance across studies and draw generalizable conclusions. This research seeks to address these challenges by systematically evaluating these tools and proposing a structured framework for their assessment.

Despite the widespread application of generative AI, no standardized framework exists for evaluating its performance in data correction. Current assessments are fragmented and inconsistent, making it difficult to compare effectiveness or establish best practices. As organizations increasingly rely on AI for data management, ensuring these tools meet reliability and scalability requirements is crucial. This study contributes to that need by developing and applying an evaluation framework to assess AI tools in structured data environments.

The methodology follows a structured five-phase framework: selection, modification, testing, evaluation, and interpretation. By reviewing AI benchmarking literature, this framework integrates established performance metrics tailored to generative AI's role in data correction tasks. Empirical experiments using the Canadian Wind Turbine Database tested the tools' performance under various error conditions. Metrics such as accuracy, precision, recall, and failure rate were used to assess their effectiveness, alongside qualitative observations of AI-generated outputs.

Findings reveal that generative AI tools perform well on simple data corrections but struggle with more complex tasks, such as identifying overlapping inconsistencies or inferring missing values from dataset structures. While ChatGPT did not exhibit frequent hallucinations or systematic duplication, it occasionally fabricated outputs instead of acknowledging task limitations, raising concerns about transparency and reliability. No performance issues were observed in processing large datasets, but AI models failed in data enhancement tasks, often producing inconsistent or incomplete results.

These findings suggest that while generative AI can reduce manual effort in basic data correction, its current limitations hinder its application in complex scenarios. Organizations should use AI tools for well-defined tasks such as error detection and simple corrections, while more intricate tasks may require fine-tuning with careful quality check. AI developers must improve transparency, ensuring models clearly indicate task failures instead of fabricating outputs. Hybrid approaches combining AI with rule-based systems or human oversight could enhance data enrichment and structured reasoning tasks. Additionally, customization options and supplementary material integration show promise for improving performance.

Although this study successfully develops and applies a generative AI evaluation framework, some limitations remain. The research primarily focuses on ChatGPT, limiting

generalizability to other AI tools. It also evaluates AI only on tabular datasets, excluding alternative data structures such as time series or partitioned datasets. The framework enforces strict input and output definitions, which may overlook AI flexibility in handling less structured formats.

Future research should expand the framework's application to other AI models, including Microsoft Copilot and Google Bard, to assess its generalizability. Additionally, testing on varied dataset types, such as time series, image, or hierarchical datasets, would evaluate the framework's adaptability. Further studies should also investigate the impact of alternative file formats (e.g., JSON, XML, Excel) on AI performance. Addressing these aspects will strengthen the framework's applicability and ensure its relevance in broader AI-driven data correction tasks.

# Acknowledgement

I would like to express my gratitude to those who have supported and guided me throughout the course of my master's thesis journey.

First and foremost, I sincerely thank my supervisor, Victor Scholten, for his continuous support, patience, and invaluable insights throughout an extended period of my master's thesis project. His guidance has been essential in shaping this research, and I truly appreciate the time and effort he has dedicated to helping me navigate this process.

I would also like to extend my thanks to Olya Kudina for her guidance and support while she was able to contribute to my research. Her feedback and perspectives greatly helped in developing the foundation of this thesis.

A special thank you to Steffen Steinert, who kindly stepped in on short notice to join my thesis committee. His willingness to engage and provide thoughtful feedback despite the time constraints has been immensely helpful in bringing this project to completion.

Lastly, I want to express my deepest appreciation to my partner, Justė Motuzaitė, for her unwavering support and encouragement over the period. Her patience, motivation, and belief in me have been a source of strength through this journey.

To all who have contributed to this thesis in one way or another - thank you. Your support has made this work possible.

# Abstract

Despite the rapid advancement of AI in various fields, there remains a gap in research concerning its suitability and evaluation for general data analysis across diverse domains. The application of AI tools to process, clean, and improve complex, incomplete, or noisy datasets has not been thoroughly explored. Key challenges include determining how effectively these tools can detect and correct errors, augment data, and ensure improvements in data quality. This research not only evaluates the performance of AI tools in these tasks but also develops a structured framework for testing their capabilities in error detection, correction, and supplementation. Using the open-source Canadian Wind Turbine Database as a case study, this study introduces intentional errors to create controlled scenarios for evaluation. The framework offers insights into the effectiveness, limitations, and best practices for applying AI tools like ChatGPT to real-world data analysis challenges.

# List of Tables, Figures, Abbreviations

**List of figures:**



**List of tables:**

# Table of contents

# 1. Introduction

## 1.1 Overview of Artificial Intelligence Development

Artificial Intelligence (AI) has undergone substantial evolution since its conceptual inception in the mid-20th century. The term "artificial intelligence" was formally introduced by John McCarthy in 1956 during the Dartmouth Conference, a pivotal event that catalyzed the establishment of AI as an academic discipline [1]–[3]. Early research attempts concentrated on symbolic reasoning and problem-solving, epitomized by the development of the General Problem Solver (GPS) by Allen Newell and Herbert A. Simon [1], [4], [5]. Progress during this foundational era was constrained by the limitations of computational resources and data scarcity [6], [7]. Nevertheless, the 1980s and 1990s marked a paradigm shift with the advent of machine learning (ML), which facilitated the development of algorithms capable of learning from empirical data without explicit programming [8]–[10]. This period also witnessed the emergence of neural networks and the refinement of the backpropagation algorithm, significantly enhancing pattern recognition capabilities [8]–[10]. The early 21st century ushered in the era of deep learning, characterized by sophisticated neural network architectures capable of processing vast datasets, thereby driving transformative advancements in domains such as speech recognition, natural language processing, and computer vision [11]–[13].

At its essence, AI encompasses the design and development of computational systems endowed with the capability to perform functions traditionally associated with human intelligence, including learning, reasoning, problem-solving, and decision-making [7], [14]. AI is broadly categorized into narrow AI and general AI [6, p.22], [15], [16]. Narrow AI, also referred to as weak AI, is engineered to execute specific tasks with high efficiency, such as virtual personal assistants, recommendation algorithms, and autonomous navigation systems [6, p.22], [15], [16]. Conversely, general AI aspires to achieve cognitive versatility, enabling systems to perform a wide array of intellectual tasks with human-like adaptability and competence [6, p.22], [15], [16]. AI models are underpinned by diverse learning approaches, encompassing supervised learning (reliant on annotated datasets), unsupervised learning (focused on uncovering latent patterns in unstructured data), reinforcement learning (driven by reward-based feedback mechanisms), and generative modeling, which facilitates the synthesis of novel data instances that mirror the statistical properties of the training data [17, p.9], [6, p.695].

A notable subset, generative AI, specializes in the autonomous creation of content, including text generation, image synthesis, and musical composition [19]–[22]. Generative AI has become increasingly influential across sectors such as healthcare, business [25]–[27], and entertainment [24], where it supports content automation, enhances creative processes, and optimizes operational workflows. Accurate and high-quality data is vital for the effective deployment of AI systems across all sectors. Data errors, including inaccuracies, inconsistencies, missing data, duplicates, and outliers, can arise at any stage of data collection, processing, or storage [53]–[55]. These issues can distort analytical models and lead to flawed insights and predictions [53], [56]–[59]. Thus, in critical applications such as healthcare, finance, marketing, or scientific research, such errors can result in financial losses, compromised patient care, legal liabilities, and damage to an organization’s reputation [53], [56]–[59]. By proactively identifying and correcting these errors, decision making

systems can operate more reliably, building trust and enabling better outcomes across diverse domains [53], [54], [60].

## 1.2 AI in data analysis

Data is at the core of modern decision-making processes, but real-world datasets often come with a range of challenges that can hinder accurate analysis [53]–[55]. These datasets are frequently incomplete, noisy, or inconsistent, making it difficult to extract valuable insights. Common problems found in datasets include missing data, incorrect values, duplicate entries, and outliers [53]–[55]. Missing data can occur when information is not recorded or lost during data collection. Inconsistent data can arise when values are entered in varying formats, such as dates or names, which complicates analysis. Noisy data refers to the inclusion of irrelevant or random information that clouds the meaningful patterns in a dataset. Outliers, which are data points significantly different from other observations, can also distort statistical analyses and lead to skewed results [53]–[55].

Traditionally, these issues have been addressed through both manual and automated correction methods [61]–[63]. Manual correction involves data cleaning by human experts who review the dataset, identify errors, and fix them accordingly. This process, while effective, is time-consuming and prone to human error, especially with large datasets [61]–[63]. Furthermore, manual correction is not scalable, making it impractical for companies and organizations that handle vast amounts of data [61]–[63]. Automated correction systems, on the other hand, use algorithms to detect and rectify data errors. These systems can identify patterns in the data to automatically fill in missing values, correct inconsistencies, and eliminate duplicate entries [61]–[63]. However, while these automated systems improve efficiency, they are often limited by predefined rules and can struggle with more complex errors, such as those that require a deeper understanding of the dataset [64]–[67].

The emergence of AI provides a more comprehensive approach to addressing these challenges. AI models, particularly those leveraging machine learning and deep learning, can analyze datasets with much greater sophistication compared to traditional methods [64]–[67]. AI systems can not only detect errors in data but also offer advanced solutions by learning from patterns within the data. For example, AI can use predictive modeling to fill in missing values based on other relevant information in the dataset, providing more accurate results than rule-based methods [68]–[70]. Additionally, AI is capable of handling large, complex datasets with multiple types of errors more efficiently than both manual and automated rule-based systems [64]–[67], [71].

A significant application of generative AI is in data augmentation, are in fields where data is scarce or expensive to collect [35], [36]. In healthcare, for instance, generative models can simulate synthetic patient data, which can be used to train machine learning models without compromising patient privacy [36], [37]. This capability is essential in medical research, where generating large and diverse datasets is crucial for developing accurate predictive models.

Despite its many benefits, generative AI also comes with challenges and drawbacks. One of the primary concerns is the ethical implications of its use, particularly in the creation of realistic yet deceptive content like deepfakes or fake news [38], [39]. The ability of AI to generate highly convincing fake content raises issues related to misinformation, security, and

privacy [38], [39]. Another challenge is the inherent bias in AI models [40], [41]. Since generative models are trained on existing data, they can reproduce and even amplify biases present in that data, leading to unfair or harmful outcomes in sensitive areas like hiring, criminal justice, or healthcare [41]–[43]. Finally, commonly used AI platforms retain the prompted text and information for continues improvement, which directly or indirectly can lead to personal or even commercial data leak [215].

Generative AI models also require significant computational resources to train, which can be expensive and environmentally unsustainable. The process of training large models often consumes vast amounts of energy, contributing to the growing concern about the carbon footprint of AI technologies [44], [45]. In addition, generative AI models often struggle with a phenomenon known as hallucination. AI hallucinations occur when models produce outputs that are plausible in appearance but factually incorrect, nonsensical, or entirely fabricated [46], [47]. This issue is particularly prevalent in natural language generation tasks, where large language models may confidently generate false statements, misrepresentations, or "hallucinated" references [48], [49]. Hallucinations can undermine trust in AI systems, particularly when used in applications like chatbots, automated customer support, or content creation, where accuracy and reliability are important [50]–[52].

Despite of this, scholars and researchers have begun to explore the potential of publicly available AI tools like ChatGPT, Microsoft Copilot, and Google's Bard for tasks such as data correction and augmentation [76]–[78]. These tools have opened up AI capabilities to a broader audience, enabling even non-experts to leverage AI for data analysis [71]. These systems can assist users in identifying data errors, suggesting corrections, and providing real-time insights based on large-scale language models [71] [191]–[195]. While these tools offer significant promise, the evaluation of their effectiveness in handling datasets techniques for deploying generative AI in data correction and wrangling have been primarily experimental, with no unified framework presented for consistent evaluation or implementation [72]–[78],[195]. Current research has focused primarily on their conversational capabilities and general usage rather than their specific utility in data analysis [79]–[81]. Furthermore, while many assessment tools are designed to keep the AI capabilities evaluated equally, they struggle to provide good assessment in the real-world conditions [82], [83]. Finally, while many attempts to deploy and evaluate commonly used LLM exist [76]–[78], the area still lacks a common approach to measure the generative AI performance [195]. This study provides a comprehensive state-of-the-art review and proposes a framework for evaluating commonly available large language models, such as ChatGPT, in the context of data error detection, cleaning, and enhancement.

## 1.3 Research gap

Despite the rapid development of artificial intelligence (AI), particularly in generative AI, its application to data analysis remains underexplored. AI models undergo rigorous testing to ensure their performance, reliability, and accuracy, using metrics such as accuracy, precision, and recall [84]–[86], alongside task-specific criteria [87], [88]. These evaluations provide benchmarks for developers and researchers to assess models in controlled environments. However, the growing versatility of AI across diverse fields complicates the selection of appropriate evaluation metrics [89].

In the context of data analysis, generative AI demonstrates the potential to manage tasks such as error detection, correction, and dataset augmentation [76]–[78]. However, its multifunctional nature complicates evaluation [86], [90], [91]. While AI tools ChatGPT excels in tasks like natural language processing, its ability to handle structured data, correct inconsistencies, and supplement missing values requires a distinct evaluation framework tailored to data-driven applications [72]–[78], [195]. Tasks, such as identifying subtle anomalies, correcting structural errors, and augmenting / supplementing incomplete datasets, often lead to unexpected or mistaken outputs, underscoring the need for more focused testing methodologies.

Generative AI shows strong potential in data analysis tasks such as error detection, correction, and dataset augmentation. Yet, its multifunctional nature presents evaluation challenges, especially when handling structured data, correcting inconsistencies, and supplementing missing values. Identifying subtle anomalies and managing incomplete datasets further underscores the need for specialized evaluation frameworks.

A preliminary literature review highlights a growing need for specialized frameworks to evaluate AI models specifically for data analytics tasks. While various testing frameworks and metrics exist for AI in general [89], [92]–[94], selecting appropriate benchmarks tailored to data correction and augmentation tasks is a significant challenge [95], [96]. As fields such as healthcare, finance, business, and education increasingly rely on accurate and clean data, it is critical to ensure AI models perform effectively in real-world, practical scenarios, not just theoretical ones [97]. Current, more traditional, non-AI methods remain human labor-intensive and require extensive fine-tuning, which can significantly limit their efficiency and scalability in addressing issues such as incorrect formatting, data duplication, or inconsistencies, ultimately impacting the reliability of data analysis.

Additionally, the literature shows a multisided understanding of how AI tools, such as ChatGPT, handle data correction tasks. While some experimental studies exist on testing AI for data correction, these studies use varied inputs, including different datasets, analysis techniques, and variables, making it difficult to compare results or draw generalized conclusions [76]–[78], [189], [190]. This lack of standardization makes it harder to create unified testing frameworks and understand AI performance in data correction and augmentation.

## 1.4 Problem statement

Building upon the research gap identified in the previous chapter, it becomes evident that while generative AI holds significant potential for data analysis tasks, critical issues hinder its effective utilization, particularly in data correction and enhancement. The rapid advancement and wide application of generative AI technologies, like ChatGPT, has outpaced the development of systematic approaches to evaluate their performance in handling data-related tasks. This lack of comprehensive evaluation methods creates uncertainty regarding the reliability and effectiveness of these tools in practical data analysis scenarios.

A primary problem is that researchers and AI users are largely uninformed about the actual capabilities and limitations of common generative AI tools, such as ChatGPT, in performing data correction and enhancement. This knowledge gap can lead to the uncritical adoption of AI tools in data-sensitive environments, potentially resulting in erroneous analyses, flawed

decision-making, and compromised data integrity. This issue is compounded by multiple smaller problems, which further exacerbate the challenges associated with the effective utilization of generative AI in data analysis.

One of the compounding issues is the increase of independent research studies and diverse methodologies, which has led to fragmented knowledge. This dispersion makes it challenging for both researchers and everyday AI users to access consolidated insights, increasing the risk of missing critical information about AI performance in data-related tasks. Furthermore, the versatility of generative AI complicates the application of a one-size-fits-all evaluation method. Different data tasks require tailored evaluation frameworks, complicating the assessment of AI performance across varied applications.

Additionally, while much research focuses on AI's creative capabilities, there is a significant gap in studies examining the practical applicability of AI tools for real-world data correction and enhancement tasks. This lack of information may lead to unforeseen issues when these tools are deployed in practical scenarios. Existing attempts to evaluate AI performance in data correction are often inconsistent and confusing. The absence of standardized evaluation protocols makes it difficult to compare findings across studies, leading to ambiguity and unreliable cross-tool comparisons.

Moreover, researchers typically employ different AI models for their studies, resulting in unclear generalizations about AI limitations. Users and researchers must navigate multiple studies to identify limitations specific to data correction and enhancement, hindering the development of a cohesive understanding.

The issues outlined in this chapter highlight the need for a structured approach to evaluating generative AI tools in data analysis. In the following chapters, these identified problems will be addressed through clearly defined research objectives and corresponding research questions. This structured approach aims to bridge the knowledge gaps, standardize evaluation methods, and enhance the practical applicability of generative AI in data correction and enhancement.

## 1.5 Research objectives and questions

### Main research objective

To address the problem and research gap outlined in the previous sections, this study has defined a series of research objectives.

**The primary objective is to develop and implement a comprehensive evaluation framework that assesses the effectiveness and limitations of widely-used generative AI tools in data analysis tasks, with a specific focus on their ability to detect, correct, and enhance data errors.**

Additionally, this research will provide a detailed evaluation of general-purpose AI tools, such as ChatGPT, Microsoft Copilot, and Google Bard, and their applicability in data correction and augmentation tasks. Through these objectives, the study will contribute to a deeper understanding of how generative AI models perform in real-world data analysis and propose improvements for future development.

**Additional research Objectives**

1. **Assess the current state of the art in general-purpose AI applications for data analysis.**
   This objective aims to review the existing applications of general-purpose AI tools in the context of data analysis. By conducting a thorough literature review and analyzing current AI tools like ChatGPT and Microsoft Copilot, this research will highlight how these tools are being used in practice and identify any gaps in their performance related to data correction, augmentation, and insights generation.
2. **Asses existing benchmarks and metrics used to evaluate the performance of generative AI tools.**
   This objective focuses on exploring the benchmarks and metrics currently applied to evaluate generative AI tools in various fields. It aims to analyze their applicability to tasks such as error detection, correction, and enhancement, and identify gaps in existing evaluation methods. By examining these benchmarks and metrics, the study aims to propose improvements or task-specific criteria that more accurately reflect the applicability and output quality of AI tools in addressing real-world data challenges.
3. **Investigate the ease of applying generative AI models to data correction and enhancement tasks within real-world datasets.**
   This objective aims to assess how effectively generative AI tools can be integrated into existing data analysis workflows to identify and rectify common data issues, thereby enhancing the overall quality and reliability of data-driven insights.
4. **Evaluate the effectiveness of generative AI models (such as ChatGPT) in handling incomplete, noisy, or inconsistent datasets.**
   The third objective will measure the performance of AI models in managing real-world datasets that are often imperfect. It will assess how well these models can identify and correct data errors, and how effectively they can provide insights despite data quality issues.
5. **Identify the technical limitations of generative AI models in data analysis and propose strategies to address them.**
   Finally, this objective will highlight the weaknesses of AI models, focusing on areas where they struggle, such as detecting subtle inconsistencies or generating accurate synthetic data in complex scenarios. By identifying these limitations, the research will propose solutions or improvements to mitigate these challenges in future applications.

## Main Research Question

**What methods can be used to evaluate the suitability of common generative AI chatbots like ChatGPT for data analysis, particularly in handling data errors?**

This question aims to explore methods for evaluating the suitability of common AI tools for data analysis, particularly in their ability to handle common data issues such as noise, missing data, and inconsistencies. The focus is on assessing the effectiveness of these tools in addressing real-world data complexities and comparing their performance to traditional data processing methods. Through this evaluation, the study seeks to provide insights into how AI tools can enhance data analysis and tackle challenges associated with imperfect datasets.

**Additional Research Questions**

1. **What is the current state of the art in the application of general-purpose AI tools for data analysis and their evaluation?**
   This research question aims to explore how general-purpose AI tools like ChatGPT, Microsoft Copilot, and Google Bard are currently being applied to data analysis. It seeks to provide a comprehensive overview of their use in data correction, augmentation, and insight generation, highlighting both their strengths and weaknesses. This question addresses a key gap in the literature, where the evaluation of these widely used AI tools in practical data analysis settings has not been deeply explored. By answering this question, the research will contribute to the ongoing discussion about the role of general AI tools in improving data analytics processes.
2. **What benchmarks and evaluation metrics are currently used to assess the performance of generative AI tools?**
   This question explores the benchmarks and metrics currently used to evaluate generative AI tools, focusing on their relevance and effectiveness in assessing tasks like error detection, correction, and enhancement. It seeks to identify gaps in existing evaluation methods and propose criteria that better align with real-world data analysis challenges.
3. **What steps are needed to prepare the selected generative AI algorithm to effectively process a dataset with noise, incompleteness, and inconsistencies?**
   This question focuses on the preparation required to apply AI to datasets with inherent issues. It will explore the preprocessing steps needed to clean and organize the data, ensuring the AI model can process it effectively. The aim is to understand how preprocessing impacts the performance of generative AI in correcting and augmenting data.
4. **How well does the selected generative AI model manage and correct errors such as noise, incompleteness, and inconsistencies in the dataset?**
   Here, the focus is on evaluating the AI model's ability to detect and correct data errors. It will analyze how effectively AI can fill in missing values, correct inconsistencies, and handle noise in datasets. By comparing AI’s performance to traditional error correction methods, the study aims to demonstrate whether AI offers a more efficient and accurate solution.
5. **What are the technical limitations of the selected generative AI model in handling datasets with errors, and how can these limitations be addressed or mitigated?**
   This question aims to identify the weaknesses of generative AI models in handling datasets with common errors. It will explore the limitations of current AI technologies, including areas where they struggle, such as detecting subtle data inconsistencies or generating accurate synthetic data in complex environments. The research will propose strategies for overcoming these limitations, offering potential improvements for future AI applications in data analysis.

By addressing these objectives and research questions, this study will provide a comprehensive analysis of how generative AI models can be applied to real-world data analysis, particularly in handling imperfect datasets. Additionally, it will offer insights into the current state of general-purpose AI tools and their potential for enhancing data analytics in various fields.

## 1.6 Research relevance

This research is highly relevant to both academia and industry due to the increasing reliance on data-driven decision-making across various sectors. Organizations often encounter challenges with imperfect datasets, such as incompleteness, noise, and inconsistencies. Generative AI models, including tools like ChatGPT, Microsoft Copilot, and Google Bard, offer promising solutions to address these issues by enhancing data analysis processes. By evaluating these tools, this study aims to improve the efficiency and accuracy of data management and analysis.

From an academic perspective, this study addresses a notable gap in existing literature by exploring the application of generative AI in dataset correction, moving beyond its traditional focus on creative tasks. It seeks to provide insights into the practical utility of these AI tools for data correction and augmentation, thereby contributing to the theoretical understanding of AI's capabilities in data analytics. Furthermore, the framework developed in this study aims to standardize AI testing and development, promoting consistency and reliability in future research and applications.

In the industrial context, this research aims to enhance data management across various sectors by evaluating generative AI models as efficient alternatives to traditional, labor-intensive methods of addressing data imperfections. As tools like ChatGPT, Microsoft Copilot, and Google Bard become more accessible, understanding their effectiveness in data correction and augmentation is crucial. Developing a standardized testing framework will inform users about common pitfalls and best practices, ensuring responsible AI deployment, minimizing errors, and maximizing data integrity.

Additionally, this study contributes to ongoing discussions around the ethical and practical implications of AI deployment. By identifying technical limitations and ethical considerations associated with generative AI tools, the research promotes more responsible and informed use of AI in practice. This includes addressing concerns related to bias, privacy, and the reliability of AI-generated outputs, which are essential for the trustworthy adoption of AI technologies.

In summary, this research aims to bridge both theoretical and practical gaps in the application of generative AI for data analysis. By offering insights into the current capabilities and limitations of AI tools in this area, the study will make valuable contributions to both academic knowledge and practical applications, helping to shape the future of AI-driven data analytics.

## 1.7 Research Scope and Clarification

This research focuses on developing a framework for testing the capabilities and limitations of generative AI models in addressing common data errors, such as noise, incompleteness, and inconsistencies. The primary objective is to propose a systematic evaluation framework that can assess the suitability of AI tools for tasks such as data correction, enhancement, and insight generation. This framework will be experimentally applied to a real-world case to validate its usability and effectiveness.

The dataset used in this study is the Canadian Wind Turbine Database, which represents typical challenges in data analysis, including incomplete or inconsistent records. To simulate real-world complexity, intentional data errors will be introduced into the dataset. The research is limited to tubular datasets (text and number-based tabular datasets), allowing the framework to focus on structured and semi-structured data. This focus reflects the frequent reliance on tabular data across domains such as business, healthcare, and scientific research.

Unlike studies that test AI tools directly, this research aims to establish a robust framework for evaluating such tools. The study will explore existing AI tools, such as ChatGPT, within the framework to demonstrate its applicability, but the primary focus remains on the framework itself rather than the AI tool's performance. This approach ensures that the framework can be applied universally across different AI tools and datasets.

The scope excludes the development of new AI models or proprietary solutions. Instead, the research focuses on creating a practical evaluation methodology that can be used by data analysts and researchers to test the performance of general-purpose AI tools in handling structured datasets. The study will not explore unstructured data tasks, such as image or audio processing, to maintain a clear focus on traditional data analysis processes.

This research contributes to the growing need for systematic approaches to evaluate AI tools for data analysis. By providing a structured evaluation framework, the study aims to offer actionable insights for optimizing AI-driven data correction and enhancement processes, ensuring relevance to real-world applications. The findings are intended to guide both the development of future AI tools and their application in addressing everyday data challenges.

## 1.8 Research Approach

This study adopts an inductive research approach, focusing on developing a framework for evaluating generative AI tools in tasks such as data correction and enhancement. The inductive approach emphasizes deriving general principles from specific observations and experimental findings, rather than testing pre-established hypotheses. Insights gained through data collection and analysis are used to structure a framework grounded in real-world applications, enabling a systematic evaluation of AI tools.

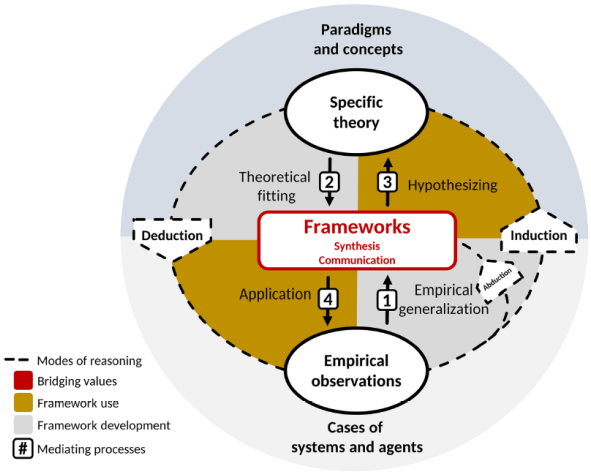



*Figure 1. A meta-framework outlining the central role frameworks [211]*

To structure the formulation and analysis of the framework, this study draws on the meta-framework proposed by Partelow [211], which describes frameworks as tools for structuring empirical and theoretical inquiry. Frameworks facilitate knowledge synthesis and communication by providing a structured methodology for organizing and analyzing research questions. Partelow's meta-framework identifies four core mechanisms for framework development and application: Empirical Generalization, Theoretical Fitting, Hypothesizing, and Application. These mechanisms often

interact and overlap, with their priority depending on the purpose and context of the framework being developed (Figure 1).

In this study, the focus is application-oriented, with the framework designed to guide real-world tasks, such as assessing AI performance in data correction and enhancement, while also generating insights into the theoretical underpinnings of AI evaluation. To achieve this, the research approach integrates multiple mechanisms from Partelow's meta-framework, adapted to the goals of this study:

Knowledge Aggregation (Literature Review): Rather than relying on empirical generalization, this study conducts a systematic literature review to establish a comprehensive, state-of-the-art understanding of AI evaluation methodologies. This approach enables the synthesis of existing knowledge, the identification of patterns, and the recognition of gaps in current AI evaluation methods. By analyzing a broad range of studies, the review forms a solid foundation for developing the proposed evaluation framework.

Theoretical Fitting and Hypothesizing: These mechanisms are used to identify and connect existing and emerging relationships between key concepts in AI evaluation. This theoretical groundwork informs the structure and components of the proposed framework, ensuring it is both conceptually robust and practical.

Framework Creation: Building on the insights from the literature review and theoretical synthesis, the study proposes a framework for evaluating generative AI tools, which is detailed in the chapter on Framework Creation.

Testing and Validation: The framework is tested and validated through a series of experiments involving AI tools and real-world datasets, such as the Canadian Wind Turbine Database. This experimental phase evaluates the framework's applicability and effectiveness in assessing AI performance under controlled conditions. Experimentation is a valuable approach as it provides empirical evidence on true AI capabilities, allowing for objective assessment and iterative refinement of the framework based on observed performance.

By integrating these methods, the research approach ensures that the proposed framework is grounded in substantial evidence, theoretically informed, and practically relevant. This multi-faceted methodology not only addresses the immediate goals of evaluating generative AI tools but also provides a structured pathway for refining and applying the framework in diverse contexts. The interaction of development and application mechanisms enhances the framework's adaptability, making it a valuable contribution to the growing field of AI evaluation.

# 2. Literature analysis

This chapter presents a comprehensive literature analysis on the role and evaluation of AI, particularly within the context of data analysis, testing, and benchmarking. The purpose of this review is to establish a theoretical foundation for the study by examining existing research, identifying key themes, and recognizing areas where current knowledge may be limited or inconsistent. Moreover, this chapter aims to build an understanding of both the technical capabilities of AI in data processing tasks and the methodologies used to evaluate AI models effectively.

### Methodology and keywords

The literature review presented in this section is structured as a thematic analysis, aimed at organizing and synthesizing relevant studies on AI's application in data analysis and the methodologies for AI testing and benchmarking. This approach allows for a focused exploration of key themes, debates, and research gaps, providing a structured foundation for the study's theoretical and methodological framework.

The research was conducted using several prominent academic databases, including Google Scholar, JSTOR, IEEE Xplore, ScienceDirect, and Semantic Scholar. These databases were chosen for their extensive repositories of peer-reviewed journals, conference proceedings, and technical papers, which ensure a comprehensive view of the current state of knowledge in AI and data analysis.

The literature search was guided by several key phrases relevant to the study's focus areas. The main research phrases included:

- "Artificial intelligence in data analysis"
- "Applications of AI in data correction and augmentation"
- "Generative AI for data processing"
- "AI testing frameworks and methodologies"
- "Benchmarking AI performance in data tasks"
- "Evaluation metrics for AI models"
- "Challenges in AI testing and validation"

These research phrases were used to identify studies that discuss both the functional application of AI in data-related tasks and the various approaches to evaluating AI effectiveness and reliability. Through this targeted search and thematic organization, the literature review seeks to capture the essential developments in the field, as well as the limitations of existing research, thereby positioning this study within the broader context of AI and data analysis.

## 2.1 Literature about AI in field

Generative AI has revolutionized data analysis by introducing advanced methods for managing complex datasets, especially in addressing incomplete or missing data. Models like

Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) are pivotal in this transformation. GANs utilize a generator-discriminator framework to produce data that closely resembles real-world distributions, while VAEs employ an encoder-decoder structure to generate new data based on learned representations [98].

A significant advantage of generative AI is its proficiency in creating synthetic data to fill gaps in datasets, enhancing completeness and accuracy [98]. Beyond generating synthetic data, AI can augment datasets by incorporating additional information through methods like web scraping [114]. This approach automates the extraction of relevant data from various online sources, enriching the dataset and providing a more comprehensive foundation for analysis. In addition to augmenting datasets, generative AI plays a crucial role in error correction. Traditional methods for error detection, such as rule-based systems and pattern enforcement, are widely used but often insufficient for complex datasets [54]. These techniques frequently require human intervention and are not well-suited for handling the intricacies of large, diverse datasets [54].

On the other hand, a recent study by [99] focused on industry showcase that generative AI offers a more advanced approach to addressing these challenges. Models like GANs and VAEs are adept at learning underlying data patterns, enabling them to identify and rectify errors in large datasets. In fields like oil and gas, machine learning techniques such as XGBoost and convolutional neural networks (CNNs) have been successfully applied to detect anomalies and correct errors in well logging and PVT (pressure, volume, temperature) datasets [99]. These models can simulate missing data and correct anomalies based on learned patterns, significantly improving data quality without requiring manual intervention [99]. This approach not only enhances the reliability of the data but also reduces the need for human oversight [100].

Generative AI has also emerged as a powerful tool for data augmentation, particularly in fields where data is scarce or difficult to collect. By creating synthetic data that closely resembles real-world conditions, AI-driven techniques can improve model performance and generalizability. For example, in healthcare, where data privacy and scarcity are major challenges, generative models like GANs and VAEs are used to generate high-quality, diverse datasets for training and testing purposes [101]. Furthermore, these models help mitigate the problem of imbalanced datasets, especially in fields like fault diagnosis and anomaly detection [102].

Augmentation can also involve merging generative AI with real data. Recent analysis by [103] on Large Language Models (LLMs) like ChatGPT shows that these models provide conversational, context-aware responses, moving away from traditional list-based search engine formats. This is key as users increasingly prefer direct answers over sifting through multiple links. While the analysis emphasizes AI-generated answers from training data, the potential of generative AI to access real-time internet information should not be overlooked. Nonetheless, despite offering efficiency and interactivity, challenges persist in academic and health fields, where accuracy, credibility, and proper source citation are crucial [104]. As the [104] study explains, generative AI models often produce responses lacking reliable references or including fabricated information, limiting their suitability for rigorous research.

Dataset enrichment enhances existing datasets by adding relevant information or improving quality for analysis. Traditional methods—such as manual integration or sourcing data from external providers—are often resource-intensive and time-consuming [105], [106].

Consequently, automating these processes has become a significant research focus [106]. Currently, techniques like web mining—extracting data from web tables or knowledge bases, crawling for supplementary information, or purchasing large datasets—are commonly employed [105]. Although effective, these methods require considerable human and computational resources, creating opportunities for AI to streamline and enhance enrichment. Multiple studies suggest that generative AI models, including large language models (LLMs), hold promise for revolutionizing data enrichment [105], [107]–[113]. These tools can automate tasks such as web crawling or scraping [105], [113]–[115], extract data from online datasets and open data sites [106], [109], and integrate diverse external sources like PDFs and scans [111]. Moreover, AI can efficiently classify and categorize data, enabling faster integration of new information into existing datasets [108], [110]. Beyond structured data, AI-enhanced methods extend to unstructured and semi-structured sources, improving datasets for AI model training with better accuracy and performance [107]. However, ethical concerns—such as data privacy and bias—must be addressed in AI-driven enrichment [116].

Beyond traditional data augmentation, enrichment and error correction, generative AI offers new insights into data analysis by enhancing decision-making processes. AI systems, when integrated into hybrid decision-management frameworks, can improve accuracy and interpretability. For instance, combining traditional decision models like Decision Model and Notation (DMN) with AI models helps identify complex, non-linear relationships between variables that would otherwise remain hidden [117], [118]. In predictive analytics, AI-driven models outperform traditional methods by uncovering deeper patterns in historical data. This is particularly valuable in fields like finance, where AI-based predictive models are used to anticipate market trends, assess risks, and improve decision-making [117], [118].

Despite of benefits, the increasing use of AI also raises concerns about bias and data privacy. AI systems often inherit biases present in their training data, leading to unfair outcomes, for example, in areas like hiring or credit scoring [119], [120]. Additionally, AI models can unintentionally introduce bias through feature selection, where attributes such as related codes or scores serve as proxies for sensitive data, skewing predictions [119], [120]. Privacy concerns are also significant, as AI systems typically gather and process large amounts of personal data, often without explicit user consent. This lack of transparency, especially in models referred to as "black boxes," makes it difficult to ensure that data is being used responsibly and securely [121].

To address the opacity of AI systems, efforts are being made to develop "explainable AI" (XAI), which seeks to make AI decision-making processes more transparent. Explainable AI helps users understand how and why AI models make certain predictions or recommendations, thus enhancing trust and accountability [122]. This is particularly important in high-stakes fields such as healthcare and finance, where decisions like approving insurance or mortgages could rely on skewed AI systems, making it crucial to ensure these decisions are ethical and transparent.

## 2.2 Literature review of AI testing and benchmarking

### Introduction to AI Testing and Benchmarking

Testing and benchmarking are crucial for ensuring the performance, reliability, and accuracy of AI models before real-world deployment [84][85]. Testing compares an AI model’s output

against known outcomes, while benchmarking compares different models using established metrics or standards [85][89]. These practices help researchers select suitable models for various applications.

As AI evolves from narrow, task-specific systems to more general-purpose tools capable of multiple tasks [6][15][16], consistent performance across diverse scenarios becomes essential. Early AI evaluations focused on simple metrics like accuracy and error rate [86]. For instance, rule-based systems and statistical models were assessed by comparing predictions with predefined rules [123]. As AI advanced, specialized benchmarks emerged for tasks such as natural language processing (NLP) and computer vision [87][88]. However, the wide range of AI tasks presents challenges in creating universal standards [89]. Different applications—like NLP, computer vision, and data analysis—require unique metrics aligned with their distinct objectives [89]. Image recognition models might be measured by precision, recall, and F1-scores [89], whereas NLP models often use BLEU or ROUGE [90][91]. This variation underscores the need for specialized benchmarks tailored to each domain.

Generative AI and general-purpose tools add another layer of complexity [89]. Generative models like ChatGPT-4 can handle text generation, image creation, and code synthesis, making them versatile but difficult to evaluate using traditional metrics [86]. Objective measures may not fully capture quality or coherence, requiring human evaluations for fluency and context [90][91]. These models can also produce errors such as hallucinations—seemingly plausible but factually incorrect outputs [124][125]. Hence, testing must consider both form and factual accuracy. Assessing general-purpose AI like as mentioned, ChatGPT, or Microsoft Copilot, and Google Bard is similarly challenging [89][86]. Because these systems can perform diverse tasks—from summarization to data analysis—a single evaluation method is insufficient [126]–[128]. Metrics differ: data analysis tasks focus on accuracy, error detection, and pattern recognition, while content generation emphasizes language or image quality [126]–[129].

Researchers have developed domain-specific benchmarks such as GLUE for NLP and ImageNet for vision [130][131]. However, these largely target narrow AI and may not fully capture the capabilities of multi-task or generative models. As AI continues to progress, developing more flexible evaluation frameworks is increasingly important, particularly for models in data analysis and other practical applications. Robust, comprehensive testing protocols remain imperative as AI systems scale, calling for sustained collaboration among researchers, industry, and stakeholders to refine these evolving practices.

## Historical Evolution of AI Testing and Benchmarking Approaches

The evolution of AI testing and benchmarking has paralleled the development of AI technologies, shifting from rule-based frameworks to advanced deep learning and generative approaches [86][123]. A historical perspective on these methodologies illuminates how current testing practices emerged and underscores the persistent challenges faced by researchers.

One early milestone was the Turing Test, introduced by Alan Turing in 1950 [132][133]. It assessed a machine’s capacity for human-like intelligence by evaluating whether a human judge could distinguish a machine’s responses from a human’s in conversation [132][133].

While groundbreaking, the Turing Test focused largely on conversational abilities and did not account for broader AI performance across multiple tasks [133][134].

Through the 1950s and 1960s, AI predominantly employed symbolic reasoning. Systems like the General Problem Solver used rule-based techniques to tackle various problems [1][4][5]. Researchers assessed these systems by simulating human problem-solving and comparing the outputs to actual human performance [135][136]. The 1980s and 1990s marked a shift toward machine learning (ML) methods, where models learned directly from data rather than solely following hardcoded rules [8]–[10]. Statistical performance metrics—such as accuracy, precision, recall, and F1-score—became standard tools [123][137]. Meanwhile, the introduction of curated datasets, such as those provided by the UCI Machine Learning Repository, enabled more systematic comparisons among ML algorithms [138][139].

As neural networks gained momentum in the late 1990s and early 2000s, AI achievements in computer vision, speech recognition, and natural language processing (NLP) necessitated more specialized benchmarks [131][140]–[143]. Challenges like ImageNet and TREC established uniform datasets and comparison criteria, driving rapid advances in domains such as object recognition and information retrieval [131][143]. Convolutional neural networks (CNNs) [144]–[146] and recurrent neural networks (RNNs) [147][148] emerged as cornerstone techniques for handling images and sequential data, respectively, spurring the creation of new evaluation methods.

In the 2010s, deep learning models reached and even surpassed human-level performance on select tasks. AlphaGo's triumph over a world champion in Go underscored the need for specialized evaluation systems in complex domains like strategic gaming [149]–[151]. Meanwhile, NLP benchmarks such as GLUE and SuperGLUE were developed to test generalization capabilities across multiple language tasks, including sentiment analysis and question answering [130][152].

Generative AI models presented further testing challenges, as conventional metrics (e.g., accuracy or precision) proved insufficient for judging creative outputs in text and images [85][98][153]–[155]. Alternative metrics like BLEU and ROUGE [156]–[158] emerged for text generation, while the Fréchet Inception Distance (FID) and Inception Score (IS) [159]–[163] gained popularity for evaluating synthesized images. These novel approaches emphasized realism and human-like attributes rather than strictly deterministic outputs.

The rise of general-purpose AI tools, including large language models, introduced additional benchmarking complexities [20][164][165]. Capable of performing diverse tasks—from code generation to content creation—they required new multi-domain benchmarks [166]. However, fixed test suites often failed to represent the full scope of these models' abilities [168]. Consequently, developing comprehensive, dynamic frameworks remains a key challenge in ensuring that AI testing and benchmarking keep pace with ever-evolving technologies.

## Current Frameworks and Standards for AI Testing

Current frameworks and standards for AI testing have been developed to address the growing complexity and diversity of AI applications [20], [164], [165]. These frameworks provide a structured approach to evaluating AI models across various domains, from computer vision

and natural language processing to generative tasks and general-purpose AI functionalities [166]. Each framework is designed for specific tasks, metrics, and use cases to provide consistent comparisons between AI models [168]. This section looks at common frameworks and benchmarks for AI testing and how they assess AI capabilities and limitations.

One of the most prominent frameworks in AI testing is ImageNet, a large visual database designed for use in visual object recognition software research, which has become a key resource for evaluating computer vision models [131]. The ImageNet Large Scale Visual Recognition Challenge (ILSVRC), introduced in 2010, involves a dataset of millions of labeled images across thousands of categories. AI models are tested on their ability to accurately classify objects within these images, with evaluation metrics such as accuracy, top-5 error rate (the proportion of images for which the correct label is not within the top 5 predicted labels), and precision [171], [131]. The ImageNet challenge has driven significant advancements in computer vision, leading to the development of powerful deep learning architectures like AlexNet, VGGNet, and ResNet, which have set new records for image classification accuracy [131], [171]–[174]. Despite its impact, ImageNet primarily focuses on object recognition and does not fully address other aspects of visual understanding, such as object detection, segmentation, or contextual scene analysis [131], [171].

In the field of natural language processing (NLP), benchmarks like GLUE (General Language Understanding Evaluation) and SuperGLUE have become popular standards for evaluating language models [130], [152]. GLUE consists of a series of language understanding tasks, including sentiment analysis, textual entailment, and question answering, that test a model's ability to generalize across different types of NLP problems [130], [152]. The performance of models is measured using task-specific metrics, such as accuracy, F1-score, or Matthew's correlation coefficient [130], [152]. SuperGLUE extends the original GLUE benchmark by including more challenging tasks, incorporating additional metrics to capture the nuances of language understanding [130], [152]. These benchmarks have been instrumental in the development of advanced language models, such as BERT, RoBERTa, and GPT, which have demonstrated significant improvements in various NLP tasks [175]–[177]. However, while GLUE and SuperGLUE measure language comprehension across multiple tasks, they do not account for conversational AI capabilities or generation quality, which are crucial for evaluating generative language models.

For generative AI tasks, traditional evaluation metrics like accuracy and precision often fall short, leading to the development of alternative methods. In text generation, metrics such as BLEU (Bilingual Evaluation Understudy) and ROUGE (Recall-Oriented Understudy for Gisting Evaluation) have been widely adopted to assess the quality of machine-generated text compared to human reference texts [156]–[158]. BLEU measures the overlap of n-grams (sequences of words) between the generated text and reference text, while ROUGE focuses on recall, considering how much of the reference content is captured by the generated output [156]–[158]. These metrics are particularly useful for evaluating tasks like machine translation and summarization, but they do not account for the coherence, factual accuracy, or creativity of the generated content [90], [156]–[158]. As a result, human evaluation is often used alongside automated metrics to provide a more comprehensive assessment of generative text models [156]–[158] [178].

In image synthesis, metrics such as the Fréchet Inception Distance (FID) [159]–[161] and the Inception Score (IS) [162], [163] are commonly used to evaluate the quality of images generated by models like Generative Adversarial Networks (GANs). FID measures the

similarity between the distribution of generated images and real images [159]–[161], while IS assesses the quality and diversity of the generated images based on how confidently a pre-trained classifier assigns labels to them [162], [163]. These metrics have been critical for benchmarking progress in generative modeling tasks. However, FID has been shown to contradict human raters, fail to reflect gradual improvements of iterative text-to-image models, and does not capture distortion levels. It also produces inconsistent results with varying sample sizes [179]. Thus, alternative metric called CMMD has been proposed [179]

For evaluating general-purpose AI tools, which are capable of performing multiple tasks across different domains, more comprehensive benchmarks have been introduced [166], [167], [169], [170]. For example, BIG-bench (Beyond the Imitation Game) is a large-scale benchmark designed to test general AI capabilities across a variety of tasks, from arithmetic and common sense reasoning to code generation and language translation [167]. The goal of BIG-bench is to assess the breadth and depth of AI models' generalization abilities rather than their performance on narrowly defined tasks [167]. While it represents a step towards evaluating general-purpose AI, the benchmark still relies on predefined tasks, which may not fully capture the open-ended capabilities of models like GPT-4 or Google's Bard [89], [180].

In addition to task-specific benchmarks, there are also cross-domain evaluation frameworks that aim to measure an AI model's ability to transfer knowledge across different tasks and domains. These include multi-task evaluation setups, where models are tested on a suite of tasks simultaneously to assess their robustness and generalization capabilities [180]. For instance, Decathlon [181], [182] in computer vision and XTREME [183] in NLP evaluate models on diverse tasks with varying data distributions to measure their ability to adapt and perform well across different settings. Such cross-domain benchmarks are particularly relevant for general-purpose AI models, which need to demonstrate versatility across multiple domains.

Despite the availability of these frameworks, there are still challenges in creating truly comprehensive and universal benchmarks for AI evaluation. The diversity of AI applications means that no single framework can cover all aspects of AI performance, leading to the need for task-specific metrics and domain-focused evaluations. Furthermore, many existing benchmarks are static [89], relying on fixed datasets and tasks that may not reflect the evolving nature of AI capabilities. However, dynamic benchmarks, while addressing this issue, might be hard to cross compare with the previous results and can be only used in the parallel comparison [89].

## Evaluation of AI for Data Correction

Evaluating AI tools for data analysis is a multi-faceted process, encompassing different perspectives and methodologies tailored to specific applications. Evaluation of, AI tools for data analysis presents unique challenges and opportunities, as these tools are designed to perform a variety of tasks across different domains rather than excelling in a single specialized area [20], [164]– [166]. General-purpose AI tools, such as ChatGPT, Microsoft Copilot, and Google Bard, are increasingly utilized for tasks such as data cleaning, augmentation, and preparation . However, the process of evaluating these tools for data analysis tasks requires benchmarks that can effectively capture their multi-functional nature while also addressing the specific requirements and challenges associated with data manipulation [184]–[187], similarly as in the study [188]. Currently, the assessment of AI

tools for data analysis focuses on three key perspectives: AI's Ability to Clean Data Through Experimental Assessment, AI's Ability to Prepare Clean Data for Training Other AI Models, and AI's Role in Assisting Humans with Data Correction

The first perspective involves directly evaluating the AI's ability to detect and correct errors in datasets [76]–[78]. This is often conducted through experimental assessments, where datasets with known errors are supplied to the AI tool, and its performance is measured based on metrics like accuracy, precision, recall, and failure rate [76]–[78]. These metrics provide insight into the AI's capability to identify and address issues such as noise, missing values, inconsistencies, and data type mismatches. Benchmarking frameworks are used to test the AI under varying conditions, such as different datasets and sizes [189], and task prompt manipulation [190]. This approach directly examines how well the AI performs in practical data correction tasks, simulating real-world applications.

Another critical aspect of evaluation involves assessing the AI's ability to clean and preprocess data for training machine learning models [191]–[195]. This evaluation focuses on how effectively the AI can enhance data quality to improve the performance of subsequent AI models. The process typically involves using the cleaned data to train a machine learning model and then comparing the performance metrics—such as accuracy, precision, and robustness—of the trained model against models trained on unprocessed or manually cleaned data [191]–[195]. This approach measures not only the AI's data-cleaning capabilities but also the downstream impact on the performance of AI systems relying on this data.

The third perspective evaluates the AI's capability to assist humans in the data correction process. This includes the AI's ability to provide suggestions, flag potential errors, and interactively collaborate with human users to refine datasets [127]. One of benchmarks evaluating LLMs performance in aiding humans is DSEval [196]. The focus of such benchmarks is on how well the AI complements human expertise, reduces manual effort, and improves overall efficiency[196]. Metrics such as task completion time, human error rates, and user satisfaction are often used to assess the effectiveness of AI-assisted workflows. This perspective highlights the role of AI as a supportive tool rather than a standalone solution.

## Evaluation of AI for data enrichment

While several studies explore the potential of AI in data enrichment [105], [107]–[113], they primarily focus on showcasing experimental results and qualitatively evaluating AI performance. None provide a systematic framework or metrics for assessing the usability and effectiveness of AI in data enrichment. This gap leaves room for developing standardized methods to evaluate AI-driven data enrichment comprehensively.

A review by Dr. MWP Maduranga and Ms. MVT Kawya [114] categorizes AI techniques for web scraping and data augmentation into distinct methodologies, such as Natural Language Processing (NLP) and Computer Vision (CV). This classification suggests that benchmarks typically used in these fields, such as BLEU for evaluating the quality of text generation in NLP or TUBench [197] for text recognition and understanding from images in CV applications, could also be applied to measure AI effectiveness in data enrichment tasks.

Beyond specific benchmarks, statistical metrics can be employed to evaluate the success of AI-driven dataset enrichment. For example, data completeness can be assessed by the

reduction in missing values before and after enrichment. Data consistency metrics, such as the proportion of resolved inconsistencies or duplicates, can indicate how well AI improves dataset reliability. Data accuracy could be evaluated by cross-checking AI-augmented data against ground truth data or external validation sources. Finally, integration success rates can measure how seamlessly AI integrates new data from external sources into existing datasets.

## Evaluation Metrics Commonly Used in AI Testing

Evaluation metrics are essential for quantifying how effectively AI models perform specific tasks, whether they involve classification, prediction, generation, or data manipulation. By providing objective and subjective assessments of model outputs, these metrics ensure that researchers and practitioners can gauge performance accurately and consistently. The selection of an appropriate metric is critical, as it highlights different aspects of a model's behavior and can influence both model development and deployment strategies.

In classification tasks, accuracy is frequently the first metric considered, reflecting the proportion of correctly identified instances [76]–[78]. However, when dealing with imbalanced datasets, accuracy may fail to capture performance on minority classes. In these cases, metrics such as precision, recall, and the F1-score offer more nuanced insights [76]–[78], [129], [198]. Precision quantifies how many of the model's positive predictions are correct, while recall measures the proportion of actual positive instances correctly identified. The F1-score balances these two, making it particularly useful in scenarios where both false positives and false negatives carry significant costs.

Regression tasks, which aim to predict continuous values, typically use metrics such as mean absolute error (MAE) and root mean squared error (RMSE) [129], [198]. MAE calculates the average absolute difference between predicted and observed values, offering a straightforward understanding of prediction accuracy. RMSE goes further by squaring these differences before averaging, emphasizing larger errors and making it more sensitive to outliers [129], [198]. Both metrics are common in forecasting and numerical modeling, where consistent predictive accuracy is paramount.

For generative AI models, especially those tasked with text generation, image synthesis, or creative content production, traditional metrics like accuracy and precision can be insufficient [129], [198], [87]. In text-based tasks, BLEU (Bilingual Evaluation Understudy) and ROUGE (Recall-Oriented Understudy for Gisting Evaluation) dominate [156]–[158]. BLEU focuses on n-gram overlap between generated text and reference text, while ROUGE assesses how much of the reference content is captured by the output. These metrics are valuable for machine translation and summarization, yet they may not fully account for coherence, creativity, or factual correctness [129], [199].

Image generation tasks often rely on the Fréchet Inception Distance (FID) and the Inception Score (IS) to evaluate realism and diversity [159]–[161], [162], [163]. FID compares the statistical distributions of generated and real images, with lower scores indicating greater similarity to real data [159]–[161]. IS measures how confidently a pre-trained classifier (e.g., Inception v3) can label the generated images and how diverse those labels are [162], [163]. While both metrics provide quantitative insights, they cannot capture subjective qualities like artistic style or visual appeal.

Comprehensive evaluations of general-purpose AI models often appear in benchmarks such as GLUE (General Language Understanding Evaluation) and SuperGLUE [130], [152]. These frameworks employ metrics like accuracy, F1-score, and Matthew's correlation coefficient to evaluate a range of tasks, including sentiment analysis and textual entailment. Although such benchmarks offer valuable snapshots of model performance, they may not fully assess conversational and generative capabilities [130], [152].

Finally, for data manipulation tasks—encompassing data cleaning, transformation, and imputation—evaluation metrics concentrate on the accuracy of changes made to the data. Mean absolute error (MAE) and root mean squared error (RMSE) remain common measures for imputation accuracy [98], [129], [198]. In data cleaning, precision and recall help quantify how effectively errors are detected and corrected [195]. Improvements in downstream model performance, such as increased accuracy or reduced error, can further indicate how successful the data manipulation processes are [191]–[195].

## Analysis of commonly used metrics

An examination of 25 articles [129], [180], [216]–[239] on AI evaluation metrics shows a variety of approaches for quantifying performance. Accuracy remains the most prevalent, appearing in 17 articles, underscoring its ongoing importance in classification tasks. Although accuracy is simple and widely applicable, it can be misleading with imbalanced datasets, as it may overlook minority classes.

Metrics like Precision, Recall (Sensitivity), and F1-Score, each mentioned in four articles, offer more nuanced insights [76]–[78]. Precision focuses on the proportion of true positives among all predicted positives, whereas Recall measures how many actual positives are correctly identified. The F1-Score balances both, making it valuable in scenarios where the costs of false positives and false negatives must be carefully managed. For generative models, Frechet Inception Distance (FID) appeared four times, illustrating its significance for assessing the realism of generated images. Perplexity, mentioned three times, measures predictive uncertainty in language models, reflecting its relevance for evaluating text generation tasks. Other metrics, including Efficiency, Robustness, and the Area Under the ROC Curve (AUC), surfaced in at least two or three articles. Efficiency gauges computational resource usage, while Robustness examines a model's resilience to adversarial inputs or data variations. AUC, crucial in binary classification, captures how effectively a model distinguishes between classes.

Metrics for numerical predictions and image quality—such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Inception Score (IS)—were referenced twice each. MAE and RMSE are standard for regression and imputation tasks, whereas IS complements FID in evaluating generated images' clarity and diversity. Finally, Diversity, Bias, and Relevance also appeared twice, highlighting concerns about fairness, variety, and meaningfulness in AI outputs. Collectively, these findings point to a growing emphasis on metrics that extend beyond mere accuracy to capture broader dimensions of AI performance.

The analysis also identified a variety of other metrics, each mentioned only once across the 25 articles, suggesting that they may be more case-specific or relevant to particular tasks. The complete list of these metrics can be found in the '*Appendix A – Less Commonly Mentioned Metrics*'. These less commonly mentioned metrics highlight the diversity of evaluation

criteria in AI testing, where different tasks require different measures of success. They reflect the need for adaptable and context-specific benchmarks, particularly as AI systems continue to evolve and are applied to new and varied domains.

## 2.3 Identification of themes, debates, and gaps

This chapter aims to explore the key themes, ongoing debates, and theoretical gaps in the field of AI evaluation. As AI continues to evolve and expand its applications, the methods used to assess its performance have grown increasingly complex and diverse. While certain evaluation practices and frameworks have become widely accepted, others are still the subject of debate or remain underdeveloped. By examining the common theories that guide current evaluation practices, contrasting different viewpoints, and identifying the areas where theoretical development is lacking, this chapter will provide an overview of the state of AI evaluation.

### Common theories

The literature on AI evaluation presents several widely accepted theories and frameworks that have shaped how the field assesses model performance across different tasks. These common theories provide the foundation for many evaluation practices and have been used to standardize AI benchmarking efforts in various domains, including natural language processing (NLP), computer vision, and data manipulation.

One of the most prevalent theories in AI evaluation is the use of task-specific metrics [130], [152], which are designed to capture the unique characteristics of different AI applications. This approach is grounded in the principle that evaluation should be closely aligned with the specific goals of a given task. For example, metrics such as accuracy, precision, recall, and F1-score are foundational for classification tasks because they measure the ability of models to correctly identify or predict categorical outcomes [89]. Similarly, in the context of generative AI, theories supporting the use of metrics like BLEU (Bilingual Evaluation Understudy) [156]–[158] for text and Fréchet Inception Distance (FID) [159]–[161] for images have been widely adopted, as these metrics provide quantitative assessments of the quality of generated content. The underlying assumption is that by using metrics tailored to specific tasks, researchers can gain a clearer understanding of a model's strengths and limitations within that domain.

Another common theory is the emphasis on quantitative evaluation through automated metrics, which allows for standardized and objective comparisons across different models. This theory is particularly evident in evaluation frameworks like GLUE (General Language Understanding Evaluation) [130], [152] for NLP and ImageNet [131] for computer vision, where standardized datasets and scoring systems enable direct performance comparisons. Quantitative evaluation is grounded in the belief that objective measures are necessary to benchmark progress and establish performance baselines in AI research. Automated metrics such as perplexity in language modeling, mean absolute error (MAE) [129], [198] in regression tasks, and area under the ROC curve (AUC) in binary classification are often used to provide consistent, reproducible results.

The literature also highlights the importance of robustness and reliability as core aspects of AI evaluation [180]. Theories in this area propose that AI models should not only perform well under ideal conditions but also maintain their performance across varying levels of data quality and environmental changes. This perspective has driven the development of robustness tests that measure how models respond to noise, adversarial inputs, or domain shifts [181], [182] [183]. Metrics assessing efficiency and system stability are used to evaluate the extent to which models can generalize beyond the specific conditions of training data, reflecting a theoretical emphasis on real-world applicability.

The evaluation of bias and fairness in AI models is another recurring theme in the literature [119], [120], driven by ethical concerns about deploying AI in sensitive areas such as healthcare, law, and finance. Theories supporting algorithmic fairness metrics suggest that AI systems should be evaluated not only for their overall accuracy but also for their performance across different demographic groups to identify any systematic disparities [214]. This theory has led to the incorporation of fairness metrics, bias detection methods, and diversity measures into AI evaluation [119], [120] to ensure that models do not disproportionately benefit or harm certain groups.

The literature also discusses theories related to explainability and interpretability, especially in areas where understanding the rationale behind model decisions is crucial. Theories in this domain suggest that AI models should not be treated as "black boxes" but rather as systems whose outputs can be explained in a way that users can understand [121].

In summary, common theories in AI evaluation revolve around task-specific metrics, quantitative assessment practices, robustness and real-world reliability, bias and fairness considerations, and the importance of explainability. These theories provide the foundation for many existing benchmarking practices and help shape the development of standardized evaluation frameworks across various AI domains. The literature indicates that while these theories are widely accepted, ongoing research continues to refine and expand them to better address the evolving capabilities and applications of AI.

## Differences in theories

The literature on AI evaluation reveals some differences in theoretical approaches, reflecting varying perspectives on how AI models should be assessed. These differences highlight ongoing debates over the appropriateness of specific evaluation methods and show the diversity of thought in the field.

One of the most prominent differences is the debate between quantitative versus qualitative evaluation metrics. Theories supporting quantitative evaluation emphasize the need for standardized, objective measures to facilitate comparison across models and studies. This approach is evident in frameworks like GLUE for NLP [130], [152] or ImageNet [131] for computer vision, where performance metrics such as accuracy, F1-score, or mean squared error provide consistent, reproducible results. In contrast, qualitative evaluation approaches advocate for the inclusion of human judgments, especially in tasks involving generative AI. For example, when evaluating AI-generated text or images, metrics such as BLEU [156]–[158] or FID [159]–[161] may not fully capture subjective qualities like creativity, coherence, or aesthetic appeal. Proponents of qualitative theories argue that human evaluations are

necessary to assess the quality of outputs that go beyond what can be measured quantitatively, though this introduces variability and subjectivity [90], [91].

The literature also presents differing views on static versus dynamic benchmarks. Static benchmarks rely on fixed datasets and evaluation criteria, allowing for consistency in longitudinal comparisons and establishing performance baselines. Theories supporting this approach argue that fixed benchmarks provide a clear target for improvement and enable researchers to track progress over time. However, there is growing support for dynamic benchmarks that evolve by incorporating new tasks, datasets, or evaluation criteria to better reflect the rapid advancements in AI capabilities. Theories favoring dynamic benchmarks suggest that the static approach can become outdated quickly [89], failing to follow the pace of increasing AI capabilities or capture emerging real-world challenges. Dynamic benchmarks, however, face criticism for potentially compromising reproducibility and making it harder to maintain consistent evaluation standards, especially over time [89].

Another key area of divergence is the debate over task-specific versus general-purpose evaluation. Traditional theories emphasize the value of task-specific benchmarks [130], [152][156]–[158], arguing that metrics should be closely aligned with the goals of particular tasks, such as classification accuracy for object detection [131] or BLEU scores [156]–[158] for machine translation. This approach allows for specialized optimization and fine-tuning of models. On the contrary, some theories advocate for the development of multi-domain or general-purpose benchmarks, such as BIG-bench [166], which test models across a range of tasks to evaluate their versatility and generalization capabilities. This debate reflects the difference between the desire for narrowly focused, high-performing models and the trend towards creating more general-purpose AI tools capable of handling a variety of tasks. Theories supporting multi-domain evaluations argue that they better reflect real-world usage, where models are often required to perform multiple tasks rather than excelling in a single domain [167].

In summary, differences in theories surrounding AI evaluation reflect varying priorities and methodologies in the field. These differences manifest in debates over quantitative versus qualitative metrics, static versus dynamic benchmarks, task-specific versus general-purpose evaluations, the approach to bias and fairness, and the emphasis on explainability. The diversity of theoretical perspectives suggests that AI evaluation is not a one-size-fits-all process, and there is a need for adaptable frameworks that can accommodate multiple approaches depending on the context and requirements of different AI applications.

## Theory gaps

The literature on AI evaluation reveals several areas where theoretical development is lacking or incomplete, pointing to some gaps that need to be addressed for more comprehensive and effective evaluation practices. These theory gaps highlight limitations in current approaches and suggest areas where further research could enhance the robustness and versatility of AI evaluation frameworks.

A significant gap in AI evaluation is the limited exploration of real-world applications. Much of the current research is conducted in controlled environments with curated datasets, which, while valuable for consistency and replicability, do not reflect the complexities AI tools face in practical scenarios. Real-world conditions often involve noisy data, incomplete

information, domain-specific nuances, and dynamic environments, all of which present unique challenges. Exploring AI tools in real-world applications or using benchmarks that simulate such conditions could uncover insights into their practical capabilities and limitations. For example, real-world tests might reveal unexpected biases, difficulties in scalability, or inefficiencies in handling unstructured or diverse data formats. These insights are critical for understanding the suitability of AI tools across domains and their ability to address unpredictable challenges.

Another significant theory gap is the lack of robust frameworks for assessing subjective qualities in generative AI tasks. Existing metrics like BLEU for text generation or FID for image synthesis capture certain aspects of output quality, but they fail to address subjective attributes such as creativity, coherence, novelty, or user satisfaction. Although human evaluations are often used to fill this gap, they introduce variability and lack standardization, making it difficult to compare results across different studies. The absence of a well-defined theoretical basis for evaluating subjective qualities limits the ability to develop automated metrics that can reliably assess these aspects. Further theoretical work is needed to establish criteria for subjective evaluation and integrate them into existing frameworks.

The evaluation of data manipulation tasks presents another area with theoretical limitations. While there are established benchmarks for classification, regression, and language understanding, there is a lack of comprehensive theories on how to benchmark tasks like data cleaning, supplementation, or transformation. Data manipulation often involves processes that indirectly affect the outcomes of downstream tasks, such as improving data quality or enhancing model training. However, current theories do not provide adequate guidance on how to measure the effectiveness of these processes, nor do they offer standardized metrics for evaluating the quality of data manipulation. There is a need for theoretical development that can connect data manipulation evaluation to the impact on downstream tasks, ensuring that benchmarks reflect the practical significance of these activities.

Another notable gap in AI evaluation is the lack of research on adapting generative AI (GenAI) tools for specific tasks. While GenAIs are highly versatile and creative, their multi-purpose design often results in suboptimal performance for specialized applications. Current research largely highlights general capabilities, with limited focus on optimizing these tools for targeted use cases. Effective adaptation requires exploring methods such as fine-tuning, prompt engineering, and integrating domain-specific knowledge. Without these efforts, GenAIs risk being treated as one-size-fits-all solutions, which may not meet the demands of specialized tasks. Addressing this gap would enhance the accuracy and reliability of GenAIs for specific applications, unlocking their full potential across various domains.

In summary, the literature highlights several critical gaps in AI evaluation theories, including the limited focus on real-world applications, the absence of robust frameworks for assessing subjective qualities, the lack of standardized benchmarks for data manipulation tasks, and insufficient research on adapting generative AI tools for specific purposes. Addressing these gaps will require further theoretical development to create comprehensive, practical, and flexible evaluation frameworks.

## Conclusion / summary

The exploration of AI evaluation highlights both established theories and critical gaps in the field. Common approaches emphasize task-specific metrics, quantitative methods, and

standardized benchmarks like GLUE and ImageNet, providing a strong foundation for evaluating AI performance across domains. Key areas of focus include robustness, bias, and explainability, reflecting the need for AI systems to be not only accurate but also reliable, fair, and interpretable.

However, significant debates persist, including the balance between quantitative and qualitative evaluation, static versus dynamic benchmarks, and task-specific versus general-purpose evaluations. These differences underscore the complexity of AI evaluation and the challenges in establishing universally accepted practices.

Gaps in the literature further highlight areas for improvement, including the lack of frameworks for assessing generative AI tools in specialized tasks, evaluating subjective qualities, and benchmarking data manipulation processes. Additionally, the limited exploration of real-world applications restricts our understanding of how AI tools perform under practical conditions.

This chapter has outlined the theoretical trends, debates, and gaps that inform the development of a more comprehensive evaluation framework. These insights will serve as the basis for addressing these challenges in the subsequent framework proposed in Chapter - 4 Conceptual Framework.

# 3. Thesis theoretical framework

## 3.1 Introduction

A theoretical framework provides the foundation for research, connecting the study's objectives, methods, and findings to established theories. It offers a lens to analyze the research problem and clarify the relationships between key concepts and variables.

In this thesis, the theoretical framework grounds the evaluation and benchmarking of AI tools, such as ChatGPT, in existing knowledge on data quality, error detection, correction, and augmentation. It situates the research within broader academic and practical contexts, ensuring systematic analysis of the tools' performance on real-world datasets.

This chapter outlines the theories, metrics, and criteria guiding the study and links them to the research question: "What methods can be used to evaluate the suitability of common AI tools for data analysis, particularly in handling data errors?" By addressing gaps in current evaluation approaches, it provides the foundation for developing and interpreting the study's framework and results.

## 3.2 Overview of Key Theories and Concepts

This section outlines the theories and concepts identified through a review of relevant literature, which provide the foundation for the Conceptual Framework. These insights help to create a foundation on how tools like ChatGPT address criteria such as accuracy, consistency, fairness, and usability in data correction and enhancement.

### Benchmarks for AI Tools and Evaluation Metrics

Clarification of the concepts of 'benchmarking' and 'metrics' is very important for this study, as they form the foundation for evaluating generative AI tools like ChatGPT [204], [205]. Benchmarking evaluates a system's performance against established standards, while metrics translate outputs into measurable values, enabling structured assessments. In artificial intelligence, benchmarks provide a way to compare models systematically, and metrics such as accuracy, precision, and F1-score assess specific aspects of performance. These tools are essential for evaluating capabilities in tasks such as data correction and augmentation.

Standard benchmarks like GLUE and SuperGLUE assess natural language understanding through tasks such as sentiment analysis and question answering, employing metrics like F1-score to measure accuracy and consistency. Similarly, ImageNet, although focused on computer vision, exemplifies structured evaluation methods that influence other AI domains. Broader benchmarks such as BIG-bench extend this approach, testing multi-task capabilities like reasoning and coding, which are particularly relevant for generative AI tools.

Metrics for generative AI extend beyond traditional measures. Precision, recall, and F1-score remain essential for evaluating tasks like error detection, while BLEU and ROUGE assess text generation. Fréchet Inception Distance evaluates generated images. However, these metrics often fail to capture subjective qualities like coherence or creativity, necessitating human evaluations to provide more nuanced assessments. Moreover, overamplifying the

importance of metrics can lead to Goodhart's law – a measure becoming a target make the measure inefficient.

Applying benchmarks to generative AI tools is challenging due to the complexity and subjectivity of outputs. Multi-dimensional results require diverse evaluation criteria, and frequent tool updates render static benchmarks less effective. Despite these challenges, benchmarks and metrics guide this study's framework, enabling systematic evaluation of AI tools for data correction and augmentation while addressing their unique complexities.

## Data Quality Dimensions and Frameworks

Data quality dimensions provide a structured way to evaluate the reliability and usability of datasets [206], [207]. These dimensions are essential for understanding the effectiveness of AI tools, such as ChatGPT, in improving data through error detection, correction, and enhancement.

Ikbal Taleb et al. [206] identify three core dimensions of data quality: Accuracy, Completeness, and Consistency. Accuracy refers to the degree to which data correctly represents the real-world phenomena it describes. Completeness measures whether all required data is present, ensuring that no critical information is missing. Consistency evaluates whether data is coherent across different sources and formats, enabling seamless integration and analysis.

| | | Data Quality Dimensions Related | | |
|---|---|---|---|---|
| | Data Quality Issues | Accuracy | Completeness | Consistency |
| Instance level | *Missing data* | X | X | |
| | *Incorrect data, Data entry errors* | X | | |
| | *Irrelevant data* | | | X |
| | *Outdated data* | X | | |
| | *Misfielded and Contradictory values* | X | X | X |
| Schema Level | *Uniqueness constrains, Functional dependency violation* | X | | |
| | *Wrong data type, poor schema design* | | | X |
| | *Lack of integrity constraints* | X | X | X |

*Figure 2. Data Quality Issues vs. Data Quality Dimensions [206]*

In addition to these primary dimensions, Fatimah Sidi et al. highlight others that contribute to data quality, like Timeliness and Safety. Although these dimensions enhance the understanding of data quality, my study places its primary focus on addressing common dataset errors—such as inaccuracies, missing data, and inconsistencies [208]—as these are the most relevant to the goals of error detection and correction.

By emphasizing Accuracy, Completeness, and Consistency, this study aligns with the fundamental attributes of high-quality data while acknowledging the importance of additional dimensions. These core concepts guide the evaluation of AI tools in ensuring that corrected or augmented datasets meet the practical requirements of reliability and usability.

## Generative AI and Large Language Models (LLMs)

Generative AI refers to a category of artificial intelligence focused on creating new content, such as text, images, or audio, based on learned patterns from large datasets. At the core of modern generative AI are Large Language Models (LLMs), which leverage advanced architectures to process and generate human-like text. These models, including tools like ChatGPT, are built upon foundational technologies such as transformers, attention mechanisms, and reinforcement learning.

Transformers are a pivotal architecture in LLMs, designed to process sequential data by focusing on relevant input elements through attention mechanisms. This allows LLMs to understand context and generate coherent outputs. Reinforcement learning further refines these models by optimizing responses based on human feedback, improving their ability to handle complex and nuanced tasks.

The capabilities of LLMs extend beyond generating coherent text. They excel in detecting patterns, identifying anomalies, and filling gaps in datasets, making them particularly relevant to data processing tasks. For example, tools like ChatGPT can suggest corrections for inconsistent or missing data, streamlining data quality improvement processes.

However, generative AI models face notable challenges. Hallucinations—instances where the model generates plausible but incorrect information—pose risks in critical applications. Bias in training data can lead to skewed or unfair outputs, while the computational intensity of training and deploying LLMs raises concerns about resource efficiency and accessibility.

Understanding these capabilities and limitations is essential for evaluating the role of LLMs in tasks such as error detection and data augmentation. This study builds on these insights to assess the effectiveness of generative AI tools in improving data quality, ensuring that their strengths are maximized while addressing potential drawbacks.

## Error Detection, Correction and Enhancement in Data Processing

Error detection, correction, and enhancement are integral components of ensuring data quality in analytical processes. Each term represents a specific aspect of addressing data quality issues, with enhancement encompassing broader efforts to improve data usability beyond merely resolving errors. This section explores these concepts, their traditional and AI-driven approaches, and the role of generative AI in advancing these tasks. It also examines the distinction between enhancement and augmentation in data processing.

Error detection involves identifying inaccuracies, inconsistencies, or anomalies in datasets. Traditional approaches rely on rule-based systems to flag deviations from expected patterns or ranges. AI-driven methods, such as anomaly detection algorithms, extend this capability by leveraging machine learning to detect subtle or complex patterns of error that might elude manual processes.

Error correction refers to resolving detected errors to restore the dataset's reliability. Traditional techniques include statistical imputation, which estimates and fills missing values based on the dataset's overall trends, and manual correction, where domain experts intervene to fix errors directly. AI tools enhance these methods by applying predictive modeling and

generative approaches to propose contextually appropriate corrections or be based on the human-like logic.

Enhancement goes beyond error detection and correction to improve the overall quality and usability of a dataset. This includes tasks such as standardizing formats, enriching data with new attributes, and increasing interpretability. Enhancement differs from augmentation, which specifically refers to generating new synthetic data points to expand the dataset, often used in machine learning to address data scarcity or imbalance. While enhancement focuses on refining existing data, augmentation introduces new data to complement it.

AI-driven approaches have revolutionized data enhancement and augmentation. For instance, generative AI tools like ChatGPT can enrich datasets by providing additional context or filling informational gaps with collected data. These tools also support augmentation by creating new data points based on learned patterns, addressing issues like class imbalance in machine learning tasks. Despite of this, the study will not investigate the synthetic data generation and dataset augmentation.

This study builds on these concepts to evaluate AI tools' effectiveness in detecting and correcting errors and enhancing datasets. While acknowledging the distinction between enhancement and augmentation, the focus of this research is on the refinement of existing data rather than the generation of synthetic data or dataset augmentation. This approach ensures that generative AI tools are assessed comprehensively for their ability to address real-world data quality challenges and improve the usability and reliability of datasets.

## Bias and Fairness in AI Outputs

Bias and fairness are critical ethical considerations in the development and deployment of AI tools. These factors significantly influence the reliability and societal impact of AI-driven data processing systems. Understanding the sources of bias and implementing frameworks to ensure fairness are essential for building trust in AI systems.

Bias in AI outputs can originate from multiple sources [209]. One primary source is the training data, which may reflect historical inequities, imbalances, or errors present in the data used to train the model. Feature selection can also introduce bias when certain attributes are prioritized over others, potentially leading to skewed outcomes. Additionally, societal biases embedded in algorithms or decision-making processes can exacerbate existing disparities, reinforcing harmful patterns or inaccuracies.

To address these challenges, frameworks for fairness-aware machine learning have been developed [210]. These frameworks focus on identifying and mitigating bias at various stages of the AI lifecycle, including data preprocessing, model training, and post-processing. Techniques such as reweighting training data, adversarial debiasing, and fairness constraints during optimization aim to ensure that AI models produce equitable outputs. However, achieving fairness often requires balancing multiple objectives, such as maintaining accuracy while reducing bias.

The practical relevance of addressing bias in AI is evident in its impact on data correction tasks. Bias in generative AI outputs can result in unequal treatment of different data segments, leading to further inconsistencies or inaccuracies in datasets. For example, errors in underrepresented data categories may be perpetuated or exacerbated if the AI tool lacks

mechanisms to account for bias. This undermines trust in AI systems and limits their applicability in sensitive domains such as healthcare, finance, and public policy.

Incorporating bias and fairness metrics into the evaluation framework is essential to ensure that AI tools are assessed not only for their technical performance but also for their ethical implications. This study integrates these considerations to evaluate the extent to which generative AI tools produce unbiased and equitable outputs, reinforcing their reliability and societal value.

## 3.3 Application of Theories

The theories and concepts discussed inform the research and shape the Conceptual Framework. By grounding the study in these theoretical foundations, we establish the methodology, metrics, and evaluation criteria.

In 'The Benchmarks for AI Tools and Evaluation Metrics,' we see a structured approach for assessing AI tools. This study emphasizes objective, quantifiable metrics—such as accuracy, precision, and recall—specifically relevant to data correction tasks. Unlike creative tasks requiring human judgment, here human evaluation is unnecessary, focusing instead on precision and reliability in handling data errors.

'The Data Quality Dimensions and Frameworks' underscore how accuracy, completeness, and consistency are crucial for error detection, correction, and enhancement. These dimensions ensure AI tools effectively improve data usability and reliability.

'Generative AI and Large Language Models (LLMs)' highlights why accuracy, completeness, and consistency matter in evaluating models like ChatGPT. Mechanisms like transformer architectures and reinforcement learning enable coherent outputs and handle data inconsistencies—vital for improving data quality. The literature also raises questions about whether these AI tools can meet the technical requirements for data analysis and correction, differentiating them from existing data wrangling systems.

'Error Detection, Correction, and Enhancement in Data Processing' shows how these three steps interrelate yet require separate AI functionalities and metrics. Error detection identifies anomalies, correction resolves them, and enhancement elevates overall data quality.

Finally, 'Bias and Fairness in AI Outputs' underscores ethical considerations. Even accurate results can carry hidden biases, leading to skewed outcomes. Addressing bias and fairness is essential for real-world deployment, ensuring that generative AI tools produce equitable results. By applying these concepts, the study builds a robust framework for evaluating how generative AI tools tackle data quality issues, integrating both technical and ethical considerations.

# 4. Conceptual Framework

This study introduces a scientific framework aimed at evaluating the performance of generative AI tools in data correction and enhancement tasks. The conceptual framework, grounded in a Theoretical framework, provides a structured methodology for assessing AI tools through a combination of qualitative and quantitative methods. It seeks to capture key dimensions of AI performance, offering insights into their strengths, limitations, and adaptability.

## 4.1 Application of the framework

The created framework consists of five distinct phases: Selection, Modification, Testing, Evaluation, and Interpretation.

**Selection Phase**
During this phase, the researcher selects the appropriate AI tools and a dataset that will serve as the basis for testing. If the analysis results are intended to be shared publicly, it is important that both the AI tool and the dataset are freely available. This ensures transparency and accessibility for other researchers.

**Time and Version Stamping**
Although not a distinct phase, it is crucial for the researcher to document the time and version of both the AI tool and the dataset. Ideally, both should be retrieved and stored to ensure reproducibility. This step enables the study to be replicated, facilitates cross-comparisons with other tools, and allows for longitudinal analysis when the same tool is updated or improved over time.

**Modification Phase**
This phase involves modifications made to the dataset's format or content. The researcher must specify which dataset format is used in the study, as different formats may not be accepted by the AI tool or could lead to variations in interpretation. Additionally, in this phase, the researcher manually introduces errors into the dataset, ensuring the errors are of appropriate amounts and difficulty levels to test the AI tool effectively.

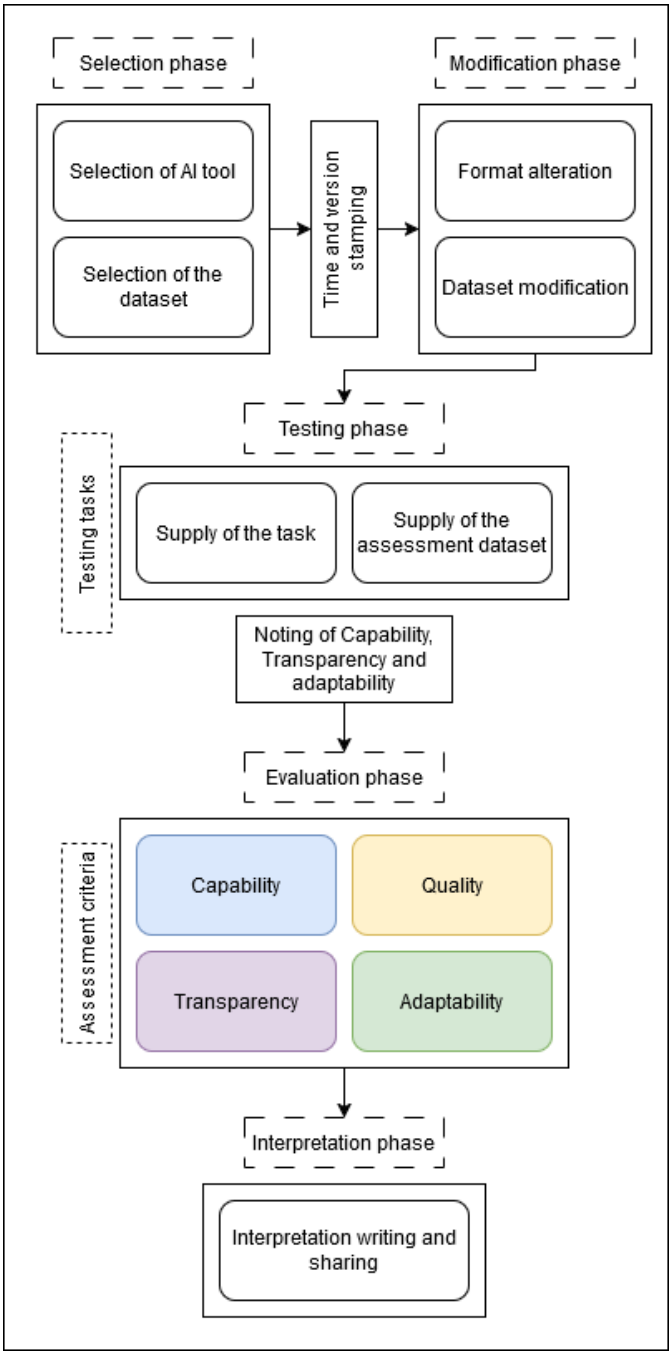



*Figure 3. Conceptual Framework*

**Testing Phase**
The testing phase includes defining and supplying the task and the assessment dataset to the AI tool. The task definition should include clear instructions on how to handle the dataset and specify the expected result format. The AI tool is expected to process the dataset and generate a corrected version in the specified file format, which can then be downloaded for further analysis.

**Noting Capability, Transparency, and Adaptability**
During the testing phase, the researcher evaluates the AI tool's performance by observing its progress, noting any errors that occur, and identifying any modifications required to execute the task properly. These observations provide insight into the tool's capability, transparency, and adaptability.

**Evaluation Phase**
In the evaluation phase, the researcher integrates the AI-corrected dataset into an analysis tool for qualitative assessment. The tool compares the original dataset, the manually adjusted dataset, and the AI-corrected dataset to produce metrics that indicate performance. The capability, transparency, and adaptability of the AI tool are also evaluated qualitatively, with detailed notes recorded.

**Interpretation Phase**
The final phase consolidates the findings from the evaluation phase. The assessment of the four criteria provides a comprehensive understanding of the tool's capabilities. The interpretation, along with detailed assessment notes and associated files, should be shared publicly alongside the published results to ensure transparency and facilitate further research.

## 4.2 Assessment criteria

The framework is designed to evaluate AI tools across four distinct assessment criteria: Capability, Quality, Transparency, and Adaptability. Each of these dimensions represents a critical aspect of AI performance and ensures a holistic assessment that goes beyond simple metrics. By structuring the evaluation into these subsections, the framework facilitates a detailed and systematic analysis of how generative AI tools perform in diverse real-world scenarios.

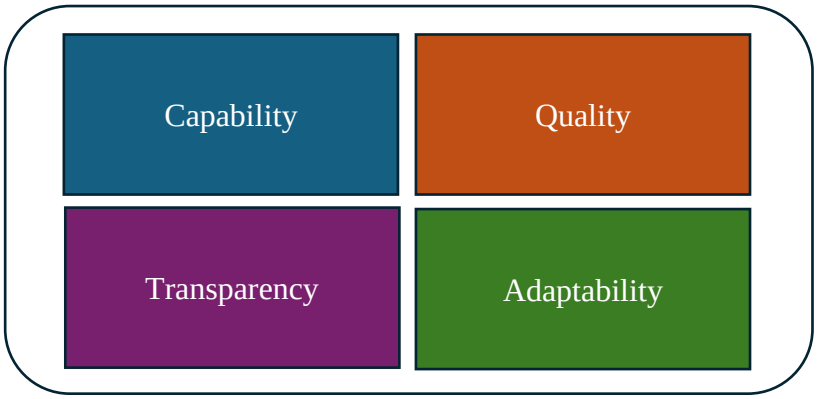



*Figure 4. Assessment criteria*

## Capability

The Capability section of the framework evaluates an AI tool's ability to analyze, manipulate, and understand datasets. This qualitative assessment focuses on key functionalities required for effective data correction and enhancement. The framework uses the following assessment to measure an AI tool’s capability:

**Data Reading Capability**
This metric assesses whether the AI tool can read and process data in different formats. While there is no universal agreement on the best formats for data analysis, tools should handle commonly used formats like .csv, .json, and .txt. The ability to accommodate a wider variety of formats increases flexibility and usability, making the tool more versatile. Tests can include supplying datasets in various formats to evaluate compatibility and functionality.

**Data Size Handling**
AI tools often have limitations regarding the size of data they can process. This metric examines the tool’s ability to handle datasets of varying sizes, typically measured in bytes, megabytes, or gigabytes. Beyond evaluating maximum dataset size, the tool’s ability to analyze subsets of larger datasets is also assessed. Testing involves determining the tool’s size limits and observing its performance with increasingly large datasets or fractions of datasets.

**Data Manipulation Capability**
Effective AI tools must not only read data but also manipulate and update it as required. This can involve tasks like correcting errors, reformatting data, or applying specified changes. AI can achieve this through two main approaches: using language model reasoning to rewrite data or generating code to apply systematic modifications. Testing involves providing specific instructions for data changes and evaluating whether the tool executes them accurately.

**Recognition of Relationships**
Datasets often contain interrelated data points where changes to one value should reflect in related fields. This metric evaluates the AI tool's ability to recognize and maintain such relationships during data manipulation. Testing involves using datasets with known interrelations and assessing whether the tool preserves these dependencies after modifications.

## Quality

The Quality section of the framework focuses on assessing the results produced by AI tools in tasks such as data correction and enhancement. This part employs quantitative metrics to measure the accuracy, reliability, and effectiveness of the AI’s output. The evaluation emphasizes the ability of the AI to recognize, correct, and augment datasets while maintaining the integrity of the original data. The following key metrics are used:

**Accuracy**
Accuracy of Error Correction is the primary metric used to evaluate the AI tool's success in detecting errors within datasets. This metric is calculated as the ratio of correctly detected errors to the total number of datapoints initially present in the dataset. Mathematically, it is expressed as:

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{True\ Positives + True\ Negatives}}{\boldsymbol{All\ data\ points}} \times \mathbf{100}\%$$

For example, if a dataset contains 100 points and 50 errors among them and the AI tool successfully detects 40 of them and 50 as non-errors, the accuracy score would be 90%. This calculation allows for a straightforward assessment of the tool's effectiveness in error detection. High accuracy in error detection indicates that the AI tool can reliably recognize data quality, making it a key indicator of the tool's overall success.

**Precision**
Precision measures the AI's selectivity in targeting errors, defined as the ratio of true positives (correctly corrected errors) to the total number of positive predictions (true positives and false positives). It reflects the AI's ability to avoid unnecessary changes to data points that are already correct.

$$\boldsymbol{Precision} = \frac{\boldsymbol{True\ Positives}}{\boldsymbol{True\ Positives + False\ Positives}} \times \mathbf{100}\%$$

In this context, "True Positives" represent data points the AI correctly identifies as erroneous and successfully corrects, while "False Positives" represent data points that the AI mistakenly corrects even though they were not incorrect. High precision means that the AI is selective, making corrections only when necessary, which reduces the risk of introducing new errors into the dataset.

**Recall**
Recall assesses the comprehensiveness of the AI in identifying errors, representing the proportion of actual errors that were successfully detected and corrected.

$$\boldsymbol{Recall} = \frac{\boldsymbol{True\ Positives}}{\boldsymbol{True\ Positives + False\ Negatives}} \times \mathbf{100}\%$$

Here, "False Negatives" refer to actual errors that the AI fails to detect and correct. High recall indicates that the AI is comprehensive in its error correction, capturing a larger proportion of the dataset's existing errors. Together, precision and recall provide a balanced evaluation: high precision ensures that the AI does not incorrectly alter data points, while high recall ensures that it captures and corrects all errors present.

**Failure Rate**
Failure Rate in data enhancement and correction is a final metric that evaluates the frequency of failure events - instances where the AI tool fails to produce an acceptable result—across different input conditions and datasets. This metric is calculated as the proportion of errors or augmentations that do not meet predefined quality thresholds relative to the total number of attempts.

$$\textbf{\textit{Failure Rate}} = \frac{\textbf{\textit{Failed Outputs}}}{\textbf{\textit{Total Outputs Attempted}}} \times \mathbf{100}\%$$

A low failure rate indicates that the AI tool can maintain a high level of quality across varied scenarios, even when faced with challenging inputs. This reliability measure is essential for assessing whether the tool is robust enough for deployment in situations where high-quality outputs are consistently required.

## Transparency

The Transparency section of the framework evaluates how well the AI tool communicates its decision-making processes and outputs. Transparency is crucial for understanding, trust, and accountability in AI-driven data correction and enhancement tasks. This section employs qualitative measures to assess the AI tool's in three parts - explainability, interpretability, and traceability.

### Explainability

Explainability measures how clearly the AI tool can articulate the reasoning behind its actions and outputs. It seeks to answer the question: How did the tool arrive at this result? This includes whether the AI provides detailed documentation or descriptions of its processes, either as it progresses through a task or when queried. Testing involves examining the AI's responses to specific questions about its decision-making process and evaluating the clarity and completeness of its explanations.

### Interpretability

Interpretability focuses on the AI's ability to justify its decisions in a way that humans can understand. It addresses the question: Why did the tool make this decision? Evaluations consider whether the AI can provide reasoning that aligns with logical or mathematical principles. Testing involves analyzing the AI's ability to explain its decisions when prompted, ensuring that users can comprehend the basis for its actions.

### Traceability

Traceability evaluates the AI tool's ability to identify and attribute the sources of its outputs. This is particularly important in tasks such as data enhancement, where external information may be incorporated into the dataset. Traceability asks: Where did the tool retrieve this information? The AI is assessed on its ability to cite sources, either automatically or when asked, ensuring that its outputs can be verified against credible references.

## Adaptability

The Adaptability section of the framework evaluates the extent to which an AI tool can be adjusted to meet specific requirements for data correction and enhancement tasks. While generative AI tools are often optimized for creative applications, data wrangling tasks pose unique challenges due to the lack of task-specific training. This gap can often be mitigated through detailed instructions or supplemental contextual information.

Adaptability is assessed qualitatively and focuses on how well the AI tool can be customized or fine-tuned to fit the specific needs of a task. Two key areas are evaluated:

**Instruction Adaptation**
This metric assesses the AI tool's ability to interpret and respond effectively to various types of instructions. It examines how precisely the tool can execute tasks based on the clarity and specificity of the provided instructions. Evaluation involves supplying the AI with different levels of instruction detail and observing its responsiveness and accuracy in adapting to the task.

**General Settings**
General settings refer to other modifiable parameters or configurations that can influence the AI tool’s performance in a specific context. This includes the ability to adjust formatting preferences, manage input-output structures, or apply domain-specific settings. Testing evaluates how these settings impact the AI's functionality and its ability to tailor its performance to the specific needs of the task.

# 5. Research methodology

This chapter outlines the methodological framework guiding this study, covering the research approach, data collection, sampling, and analysis techniques employed to address the research questions. Given the study's focus on objectively evaluating AI tools, a primarily quantitative approach is taken, with some human-assisted evaluations for subjective assessments.

The methodology includes the selection criteria for datasets and AI tools, ensuring a diverse representation of data complexities and common errors. Through an experimental design, AI tools are evaluated systematically across various metrics—such as accuracy, robustness, and fairness—to provide a comprehensive view of their performance.

Each section in this chapter lays out a structured path for data preparation, analysis, and evaluation, establishing a reliable basis for interpreting the study's findings and ensuring replicable results.

## 5.1 Research Approach

The research approach adopted in this study is primarily qualitative, while retaining some quantitative traits to test the framework and draw conclusions about the current state of AI tools and their suitability. By focusing on specific metrics such as accuracy, consistency, and robustness, the study aligns with a scientific, data-driven framework that enables precise, replicable evaluations of AI performance within a predominantly qualitative context.

While this study emphasizes qualitative analysis to explore the broader implications and suitability of AI tools, it incorporates quantitative measurements to provide objective assessments across defined criteria. AI tool performance is evaluated based on metrics such as error correction, data augmentation accuracy, and robustness to varying input conditions. This structured evaluation offers a clear basis for assessing AI tools, while limited human-assisted evaluations are used to capture qualitative aspects—such as creativity and relevance—that automated metrics may not fully address. This mixed-methods approach ensures a comprehensive assessment by balancing objective measurements with subjective human judgment where necessary.

## 5.2 Research Strategy and Time Horizon

The research strategy adopted for this study is experimental, focusing on systematic testing phases designed to evaluate AI tools based on specific performance metrics. By conducting controlled tests, this approach enables a thorough assessment of each tool's capabilities in data correction and augmentation under varying levels of difficulty and complexity. The experimental design allows for consistent, repeatable results, which are essential for the objective comparison of AI performance.

This study follows a cross-sectional time horizon, focusing on a single testing period scheduled for November 15th. This snapshot approach captures the current capabilities of AI

tools without tracking changes over time. However, the study framework is designed with adaptability in mind, allowing for similar studies in the future to evaluate progress and improvements in AI performance over the years.

The timeline includes specific phases for data preparation, systematic testing, and final analysis. Data collection and preparation occur prior to November 15th, with the testing phase following immediately after. Analysis of results and report preparation will be conducted upon completion of the tests, with findings reflecting the AI tools' current performance against established metrics. This clear timeline ensures an organized approach, while the cross-sectional design offers a relevant, timely evaluation of present-day AI capabilities.

## 5.3 How RQ will be answered

This section describes the methodological approach for each research question, outlining the tests, metrics, and analytical steps involved. The additional questions are tackled individually to examine AI performance in data correction, preprocessing requirements, error handling, and limitations. These questions focus on technical assessments, pinpointing where AI tools excel and where they face challenges in data correction and augmentation. Each question employs distinct metrics, datasets, and evaluation methods to ensure targeted insights.

The primary research question—investigating methods to evaluate the suitability of common generative AI chatbots like ChatGPT for data analysis, particularly in handling data errors—is addressed by integrating insights from both the literature review and empirical findings. This combined approach highlights how these AI tools manage real-world data, focusing on relevant metrics, frameworks, and performance indicators that assess their reliability in detecting and correcting errors. By synthesizing the evidence gathered, the study provides a rounded perspective on the strengths, limitations, and potential of generative AI chatbots in data analysis contexts.

### Methodology to Answer the Main Research Question

Question: “What methods can be used to evaluate the suitability of common generative AI chatbots like ChatGPT for data analysis, particularly in handling data errors?”

The main research question - focused on evaluating methods to assess the suitability of common AI tools for data analysis- will be addressed through a combination of an already completed literature review and empirical evaluation. This approach offers a comprehensive view of existing AI evaluation techniques and data correction methodologies, detailing their strengths, limitations, and applicability to data analysis tasks.

The literature review, covered in previous chapters, explored recent advancements in AI evaluation techniques and methodologies for data correction. It examined how these methods are applied to assess AI tools' effectiveness in handling common data issues such as noise, missing values, and inconsistencies. By focusing on these evaluation frameworks and data correction strategies, the review captured the current state of practice and provided a foundation for empirical assessment by identifying key gaps and best practices.

Following the literature review, an empirical evaluation will be conducted to further assess the practical capabilities and limitations of selected AI tools in data analysis. These tools will be tested on tasks relevant to this study, such as data correction and augmentation. The evaluation will utilize both quantitative metrics (e.g. accuracy in data correction, failure rate of augmented data) and qualitative assessments (e.g. capability and transparency) to comprehensively capture tool performance. This mix of quantitative and qualitative measures ensures a balanced evaluation, allowing for an objective comparison of each tool's effectiveness and usability in real-world applications.

By addressing the main research question through these combined methods, this section will clarify how each sub-research question contributes to the understanding of AI tools in data analysis and support the creation of an evaluation framework. Each sub-question targets a specific aspect of the tools' evaluation framework - ranging from the current state of AI evaluations to what are the technical limitations of AI tool applications for data wrangling.

## Methodology to Answer the First Research Sub-Question

Question: What is the current state of the art in the application of general-purpose AI tools for data analysis and their evaluation?

To address this sub-question—understanding the current state of generative AI in data analysis—this section combines insights from the literature review with an empirical evaluation. The literature review, detailed in earlier chapters, examined recent developments in AI evaluation metrics, benchmarking, and data correction methodologies. It highlighted key strategies, common challenges, and limitations in how AI tools are assessed for tasks like data correction and augmentation.

Building on these findings, an empirical evaluation will test selected AI tools on relevant tasks, using both quantitative (e.g. correction accuracy, consistency of augmented data) and qualitative (e.g. usability, efficiency) measures. For instance, tools may be evaluated on handling incomplete datasets, resolving inconsistencies, or improving data quality through augmentation. By focusing on these tasks, the evaluation provides practical insights into each tool’s capabilities and shortcomings.

The combined outcomes from the literature review and empirical work will clarify AI’s current potential and limitations in data analysis. This knowledge underpins subsequent sub-questions, guiding considerations like model preparation, error-handling techniques, and technical constraints. Ultimately, it establishes a cohesive framework for evaluating the suitability and effectiveness of AI tools.

## Methodology to Answer the Second Research Sub-Question

Question: What benchmarks and evaluation metrics are currently used to assess the performance of generative AI tools?

This question is addressed through an analysis of the findings from the completed literature review on AI benchmarks and evaluation metrics. The review explored established

benchmarks such as GLUE, SuperGLUE, and BIG-bench, as well as task-specific metrics like BLEU, ROUGE, precision, recall, and F1-score. It also examined domain-specific benchmarks and metrics relevant to data analysis tasks, including error detection, correction, and enhancement.

From the literature, common themes and applications were identified, highlighting the strengths and limitations of current evaluation methods. These include their effectiveness in assessing quality, accuracy, and applicability in real-world scenarios. Particular attention was given to the extent to which these benchmarks capture the unique capabilities and challenges of generative AI models, especially when dealing with noisy, incomplete, or inconsistent datasets. The insights gained from this analysis form the basis for evaluating the suitability of these benchmarks and metrics for data quality improvement tasks.

## Methodology to Answer the Third Research Sub-Question

Question: What steps are needed to prepare the selected generative AI algorithm to effectively process a dataset with noise, incompleteness, and inconsistencies?

Generative AI models must be presented with data in a consistent manner without stripping away the very imperfections they are meant to address. The aim is to preserve noise, incompleteness, and inconsistencies while offering a standardized framework that facilitates fair comparisons of the models' abilities. This approach highlights each model's natural capacity to recognize and handle errors, rather than relying on extensive data cleaning.

A clear format will be applied in three key areas: instructions, file format, and interpretation. First, the model will receive structured guidance on how to treat noisy or incomplete data without prescribing explicit corrective procedures. This allows the AI to exhibit its native strengths and weaknesses in managing data anomalies. Second, the dataset will be stored in .csv files to ensure simplicity and compatibility. Despite being standardized in layout, the data will retain its original flaws, such as missing values or contradictory entries, so that the AI must engage with these challenges directly. Third, the AI's outputs will follow a structured pattern for documenting any detected errors, along with suggestions for corrections or augmentations. By labeling errors consistently, the study can track how each model attempts to mitigate or resolve data issues.

Evaluating the impact of this standardization will involve measuring the models' performance on unaltered data, with metrics such as accuracy and recall indicating how effectively they manage real-world imperfections. This measurement is a critical test of the models' abilities and reveals whether simply formatting the data clearly—without cleansing or filtering—can enhance their performance. The results will help determine if a standardized approach alone is sufficient for improving generative AI outcomes, guiding future best practices for presenting data with minimal preprocessing while still gaining reliable insights from flawed and complex datasets.

## Methodology to Answer the Fourth Research Sub-Question

Question: How well does the selected generative AI model manage and correct errors such as noise, incompleteness, and inconsistencies in the dataset?

To address the fourth research sub-question—exploring how generative AI models process and manage datasets with noise, incompleteness, and inconsistencies—this section employs a structured testing approach to evaluate the models' capabilities in handling common data issues. This evaluation focuses on assessing each AI model's ability to detect, interpret, and correct various types of errors, providing insight into their performance in real-world data analysis tasks.

The structured testing approach involves introducing synthetic errors into datasets to represent noise, missing values, and inconsistencies at different levels of complexity. Datasets will be categorized into three difficulty levels: easy, moderate, and difficult. For instance, an "easy" dataset may contain minor noise or isolated missing values, while a "difficult" dataset could include overlapping inconsistencies requiring more sophisticated correction strategies. Using controlled levels of error complexity allows for a detailed analysis of each model's strengths and weaknesses in addressing specific data challenges.

Metrics such as accuracy, precision, recall, and failure rate will be applied to quantify each AI model's effectiveness in managing errors. Accuracy will measure the proportion of errors correctly identified and corrected. Precision will evaluate the model's ability to target genuine errors without unnecessary modifications, while recall will assess its thoroughness in capturing all errors. Failure rate will measure the proportion of errors that remain uncorrected, providing an additional indicator of the model's reliability and robustness.

The results from this testing will highlight each AI model's effectiveness and reliability in managing errors, offering a practical assessment of their capabilities for real-world data analysis applications. By understanding how well these models handle datasets with common imperfections, this section will contribute to evaluating generative AI's suitability for data analysis tasks. It will also provide insights into optimal preparation steps and help identify technical limitations addressed in subsequent research questions.

## Methodology to Answer the Fifth Research Sub-Question

Question: What are the technical limitations of the selected generative AI model in handling datasets with errors, and how can these limitations be addressed or mitigated?

To address this question, the study will analyze technical limitations of the generative AI model, such as dataset size constraints, acceptable input formats, bugs, issues, and inconsistencies observed during testing. Information will be drawn from two main sources: documentation and research regarding the AI tool (e.g. ChatGPT) and empirical observations from the testing phase. By reviewing available documentation and literature about the model, the study will identify known limitations, including maximum dataset size, format requirements, and scenarios where the model struggles to process data effectively. Testing observations will complement this information by revealing practical challenges encountered during experimentation, such as difficulties in handling large datasets, incompatibility with specific formats, or unexpected behaviors like failures to correct certain errors or introducing new inconsistencies into the output.

The limitations will be categorized based on their nature, such as input constraints, operational inefficiencies, or inconsistencies in output quality. The study will also assess

whether these issues are inherent to the AI tool’s architecture, such as limitations in processing memory due to model design, or related to external factors, such as dataset complexity or formatting. Based on these insights, recommendations will be proposed to mitigate or address these limitations. These may include adjustments to dataset preparation, exploring alternative file formats for compatibility, or implementing additional preprocessing and postprocessing steps to reduce the impact of the AI model’s constraints. By combining insights from existing sources and testing observations, this approach will provide a thorough evaluation of the AI tool’s limitations and offer practical strategies for improving its reliability and applicability in handling datasets with errors.

## 5.4 Experimental setup

This subchapter outlines the framework for testing generative AI tools in handling common data issues like noise, missing values, and inconsistencies. It aims to create a controlled yet practical setup to produce results applicable to real-world scenarios.

The discussion begins with the rationale and process for selecting datasets, emphasizing their relevance to the research objectives and highlighting the key aspects like size and complexity. It then delves into how synthetic errors are introduced into the datasets to replicate common issues encountered in data, such as missing entries or inconsistencies.

This subchapter focuses on the selected generative AI tool, explaining the rationale behind its selection based on criteria such as accessibility, versatility, and alignment with the research objectives. It discusses the tool's configuration and functionality in relation to the experimental tasks. Additionally, it addresses the limiting factors affecting its performance, such as data format compatibility, dataset size constraints, and the tool's dependence on pre-existing knowledge.

Finally, the overall experimental design is described, providing a clear procedure for conducting the tests in a consistent and reproducible manner.

### Dataset selection

#### Selection criteria

The study by Cagatay Catal and Banu Diri [189] examines how dataset size affects software fault prediction performance using five NASA datasets ranging from 498 to 10,885 modules. They found that larger datasets, benefit significantly from complex algorithms such as Random Forests, which deliver the best prediction performance in terms of AUC. In contrast, smaller datasets are better suited to simpler algorithms like Naive Bayes. The findings highlight that dataset size strongly influences algorithm effectiveness, with larger datasets amplifying the performance advantages of more sophisticated models. Overall, the study underscores the need to align algorithm choice with dataset size to optimize fault prediction outcomes.

Datasets with fewer than 10,000 records (data points) are generally manageable for most algorithms, even on modest hardware. Datasets exceeding 100,000 records or having high dimensionality (e.g. hundreds or thousands of features) can strain memory, storage, and processing power, particularly on personal computers or non-distributed systems. According

to the official ChatGPT support site [212], the maximum file size limit is 512 MB per file. For CSV files or spreadsheets, this typically translates to a practical file size limit of approximately 50 MB due to additional memory requirements for parsing and processing the data.

The dataset for this study must reflect real-world data challenges, offering a meaningful context for evaluating the generative AI tool's capabilities. It should exhibit common issues like noise, missing data, and inconsistencies, enabling a robust assessment of how well the tool can address these problems.

Real-world applicability is critical; the dataset should represent scenarios encountered in practical data analysis tasks. While synthetic errors may be introduced to simulate specific challenges, the dataset should primarily reflect realistic conditions to ensure the study's findings are relevant and transferable.

Data complexity is another essential factor. The dataset should include a mix of simple and complex attributes, such as inconsistent formats, duplicate entries, and diverse data types like numerical, categorical, or textual data. This variety tests the tool's flexibility and performance across different data challenges.

The size of the dataset should be carefully considered. It must be large enough to provide meaningful insights into the tool's scalability while remaining manageable within computational constraints. This balance ensures the dataset is neither too trivial nor excessively burdensome, allowing for a realistic evaluation.

Structure and diversity are also important. The dataset may consist of structured, semi-structured, or unstructured data, depending on the scope of the evaluation. A diverse dataset ensures the tool's adaptability to various data forms, providing a comprehensive view of its functionality.

Accessibility and licensing must also be addressed. The dataset should be publicly available or accessible within legal and ethical guidelines, ensuring transparency and reproducibility. Additionally, it is beneficial for the dataset to include annotations or a well-defined ground truth for benchmarking corrections or augmentations.

Finally, domain specificity can add depth to the evaluation. If the study focuses on a particular field, such as healthcare or finance, the dataset should reflect typical challenges from that domain, enhancing the study's contextual relevance.

**Dataset selection**

The Canadian Wind Turbine Database [213] offers detailed information on wind turbines installed across Canada, including geographic locations and key technological specifications. This dataset was compiled collaboratively by CanmetENERGY-Ottawa, the Centre for Applied Business Research in Energy and the Environment at the University of Alberta, and the Department of Civil & Mineral Engineering at the University of Toronto, under the oversight of Natural Resources Canada. It is important to note that total project capacity figures are derived from publicly available sources and may not align exactly with the sum of individual turbine capacities due to factors such as de-rating. The database is regularly updated, and users are encouraged to report errors or provide additional information via the

contact email provided on the dataset's page. Dataset record details: Released on 2020-06-19, last modified on 2024-10-08, Record ID: 79fdad93-9025-49ad-ba16-c26d718cc070.

The dataset was retrieved on 2024-11-20 in Excel format. It consists of 7578 rows and 18 columns, including attributes such as: Province_Territory, Project Name, Total Project Capacity (MW), Turbine Rated Capacity (kW), Rotor Diameter (m), Hub Height (m), Manufacturer, Model Commissioning, Latitude, Longitude, and others. In total, the original dataset contains 136,404 data points, combining simple attributes like turbine capacity with complex ones such as geographic coordinates and project-level discrepancies. This mix offers a diverse challenge for AI tools, allowing an evaluation of their ability to handle numerical and categorical data while addressing inconsistencies in formats and values.

With over 136,000 data points, the dataset exceeds the 100,000-point threshold, providing ample data to test scalability. Despite its manageable Excel file size of 811 KB—well below the 512 MB file size limit—the dataset's complexity and structure may still challenge some AI tools. However, processing the entire dataset would exceed practical time constraints without an efficient method for introducing and tracking errors. To maintain feasibility within the study's timeframe, a subset of the dataset was selected, preserving its diversity and real-world applicability.

The dataset's public availability under Canada's Open Government License ensures ethical and legal compliance, supporting transparency and reproducibility in AI research. Furthermore, its focus on renewable energy introduces domain-specific challenges, such as capacity discrepancies and geographic variations, which are common in the energy sector. This makes it a relevant and valuable resource for evaluating the adaptability and performance of the AI tool in addressing diverse and practical data challenges.

**Dataset sizing**

The original dataset was initially sorted alphabetically by the "Province_Territory" column, followed by "Province_Territoire," then "Project Name," and finally numerically by the "Turbine Identifier," which also includes the Turbine Number. If the top rows of the dataset were selected to reduce its size, the resulting subset would be imbalanced and fail to capture the diversity of data points present in the full dataset. To address this issue, a more representative sampling method was required.

A target size of 75 rows, or 1350 data points (75 rows × 18 columns), was determined to be manageable while still preserving diversity. To achieve this, systematic sampling was applied in two steps, with every 10th row of the original dataset selected in each step. This process initially produced a dataset of 76 rows. To meet the 75-row target, the last row (76th) was removed.

To assess the AI tool's performance across different dataset sizes, the 75-row dataset was further divided into three subsets of varying sizes: 25 rows, 50 rows, and 75 rows (picture). For better tracking and analysis, an additional column was added to indicate the original line numbers from the dataset. This modification increased the total data points for the 75-row dataset to 1425, accounting for the additional column.

No other changes or modifications were made to the dataset, ensuring that it retained its original attributes and structure while enabling a systematic evaluation of the AI tool's capabilities across different scales.

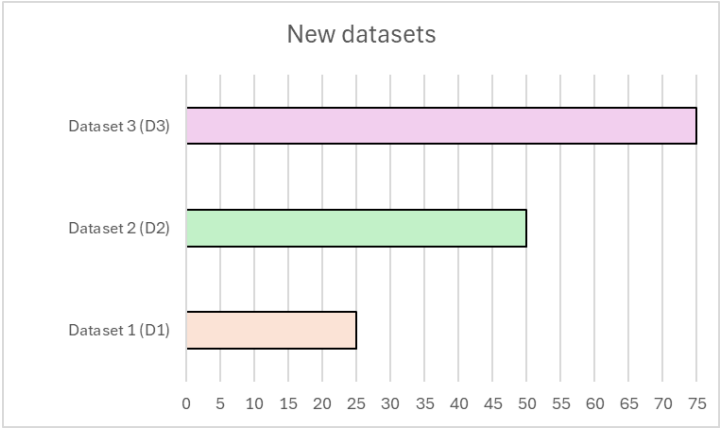



*Figure 5. Dataset division*

## Introduction of dataset errors for Experiment 1

To thoroughly evaluate the AI tool's capacity to address a wide array of data inaccuracies, errors were categorized into three levels of difficulty based on their inherent complexity. A detailed explanation of these levels, along with real-world examples, is provided in *Appendix C – Data Difficulty Evaluation Criteria*'.

**Easy Errors:**

Data Type Errors, Data Entry Errors, and Duplicate Data Errors are considered straightforward to identify and resolve due to their simplicity and the availability of automated tools. Data Type Errors, such as text in numeric fields or invalid dates, can be detected using validation tools that compare entries against predefined formats or schemas. Similarly, Data Entry Errors, such as typos or missing values, are easily flagged using pattern recognition and corrected with standard methods like validation rules or lookup tables. Duplicate Data Errors are typically resolved through matching algorithms that identify and consolidate duplicates based on well-defined criteria. These errors require minimal context or domain expertise and are highly suited for automation.

**Moderate Errors:**

Structural Errors, Inconsistent Data, and Incorrect Data Values are categorized as medium complexity because they demand more nuanced analysis and often involve patterns or relationships within the data. Structural Errors, such as misaligned columns or missing headers, require an understanding of the dataset's organization to identify deviations from expected formats. Inconsistent Data, such as varied formats, naming conventions, or units, necessitates normalization and sometimes contextual knowledge to standardize effectively. Incorrect Data Values, like out-of-range entries or logical inconsistencies, require rules and context to detect and correct, as these errors may appear valid initially. Resolving these errors

is more challenging because they often span multiple data points and may depend on domain-specific rules.

**Difficult Errors:**

Missing Data, Outliers and Anomalies, and Data Integrity Violations are considered the most complex because they often require significant context, domain expertise, and sophisticated methods to address. Missing Data involves decisions about whether to impute, interpolate, or remove values, as these choices affect the validity of analyses. Outliers and Anomalies need careful evaluation to determine whether they represent valid extreme cases, errors, or rare but meaningful events, often using advanced statistical or machine learning techniques. Data Integrity Violations, such as broken relationships or logical inconsistencies, demand an understanding of the dataset's structure, constraints, and intended purpose to resolve effectively without introducing additional errors. These errors are both harder to identify and more consequential, as mishandling them can compromise the reliability of analyses.

*Table 1. Data difficulty distribution*

| Rank | Data Error Type | Detection Difficulty | Resolution Difficulty | Overall Difficulty Level |
|---|---|---|---|---|
| 1 | Data Type Errors | Easy | Moderate | Easy to Moderate |
| 2 | Data Entry Errors | Easy to Moderate | Moderate | Moderate |
| 3 | Duplicate Data | Moderate | Moderate | Moderate |
| 4 | Structural Errors | Moderate | Moderate to Difficult | Moderate to Difficult |
| 5 | Inconsistent Data | Moderate | Difficult | Moderate to Difficult |
| 6 | Incorrect Data Values | Moderate | Difficult | Moderate to Difficult |
| 7 | Missing Data | Moderate | Difficult | Moderate to Difficult |
| 8 | Outliers and Anomalies | Difficult | Difficult | Difficult |
| 9 | Data Integrity Violations | Difficult | Difficult | Difficult |

To test the AI tool's performance across these error types, each difficulty level will introduce 5 errors per 25 rows of the dataset. Errors will not accumulate across difficulty levels and will be introduced to a sampled version of the dataset. This ensures that each level of error complexity is tested independently, allowing for a clear assessment of the AI tool's capability to address each category.

*Table 2. Data difficulty per dataset size*

| | | Difficulty level | | |
|---|---|---|---|---|
| | | **1** | **2** | **3** |
| Dataset size | **25** | 5 | 5 | 5 |
| | **50** | 10 | 10 | 10 |
| | **75** | 15 | 15 | 15 |

## Introduction of dataset gaps for Experiment 2 – Data enhancement

The full sampled 75-row dataset will be used for Experiment 2, which is divided into three parts: 2.1 Extracting data from web tables or knowledge bases, 2.2 Crawling web sources for additional information, and 2.3 Combining the extracted data with existing datasets. To simulate real-world data gaps, information will be removed from three specific columns: [Province_Territory], [Commissioning], and [Total Project Capacity (MW)], while leaving the column labels intact. These categories were chosen because their information can be retrieved from all three channels: databases, web sources, and wind turbine project websites. Removing data from these columns ensures that the AI tool can effectively locate and retrieve the missing values. Columns with data that cannot be reliably retrieved through external sources, such as [Notes], were excluded to maintain the experiment's focus and practicality.

For part 2.3, combining the extracted data with existing datasets, the experiment will include a second dataset containing the removed data points along with additional columns, such as [Project Name] and [Turbine Number], presented in a mixed order. This setup tests the AI tool's ability to align and integrate data accurately from external sources.

## Task Definition for the Experiment 1 – Data correction

One important success factor in receiving accurate and expected results from an AI tool is the quality of the task input provided [200]–[202]. The process of providing input or instructions to an AI model, like ChatGPT, to guide its behavior and responses is commonly referred to as prompting, while the scientific field dedicated to designing, refining, and optimizing prompts for the best possible output is known as prompt engineering.

Designing effective prompts is essential for obtaining the desired output from generative AI models. This process involves a range of techniques, from straightforward to more advanced approaches. For example, zero-shot prompting relies on no examples, while few-shot prompting guides the model using one or more illustrative examples [203]. More advanced techniques, such as chain-of-thought prompting, tree of thoughts, and directional stimulus prompting, can significantly enhance reasoning and improve the quality of the AI's output.

However, this study will not focus on any particular prompting technique but rather on the attributes these techniques have in common and the principles of well-designed prompting practices [203]. These include defining clear objectives and background, guiding complex reasoning, incorporating examples, and encouraging multistep interactions. The detailed prompting scripts can be found in the '*Appendix B – Prompting scripts*'.

### Definition criteria

To evaluate how different aspects of prompt design influence the AI tool's performance in dataset correction tasks, prompting will be systematically varied across four key categories: defining background and clear objectives, complex reasoning, incorporating examples, and encouraging multistep interactions. For each category, prompts will be adjusted across three levels of complexity to assess their impact on the AI's ability to identify and resolve dataset errors. Below, it outlined the specific adjustments for each category.

**Defining background and clear objective**

The clarity and detail of background information and task objectives provided will be adjusted with each level:

- Level 1 - Detailed Context: Prompts will provide a comprehensive description of the dataset's general structure, potential data types, and expected outcomes (e.g. "This is a tabular dataset containing numeric, categorical, and date fields. Ensure consistent date formats, handle missing values appropriately, resolve any formatting inconsistencies, and identify outliers where applicable. Maintain logical consistency across fields while cleaning the data").
- Level 2 - Moderate Context: Prompts will describe the dataset and highlight specific corrections required (e.g. "This is tabular dataset. Identify and correct missing values and format inconsistencies").
- Level 3 - Minimal Context: Prompts will include simple or incomplete instructions (e.g. "Fix the errors in this dataset").

**Complex reasoning**

The complexity of reasoning required to perform the task will be progressively increased:

- Level 1 - Advanced Reasoning: Prompts will require the application of domain-specific logic or handling more intricate relationships across fields. For example: ("Make a check if no interdependencies exist between datapoints and make corrections accordingly" and "Ensure that numeric fields match logical constraints").
- Level 2 - Moderate Reasoning: Prompts will include tasks requiring logical relationships or dependencies between fields. For example: ("Make a check if no interdependencies exist between datapoints and make corrections accordingly").
- Basic Reasoning: Prompts will involve straightforward tasks without interdependencies.

By decreasing the reasoning complexity, this category will assess the tool's ability to handle interdependent data corrections and logical constraints.

**Incorporating examples**

The inclusion of illustrative examples in the prompt will be varied to assess their impact on the AI's ability to perform dataset correction tasks:

- Level 1 - Multiple Examples: Prompts will provide 3 illustrative examples for each error type in the ranking, addressing different types of common errors.
- Level 2 - Single Example: Prompts will include a single generic illustrative example of a specific correction.
- Level 3 - No Examples: Prompts will include only general instructions, without any illustrative cases.

These adjustments will allow the study to measure how examples influence the AI's understanding of error correction tasks and its ability to generalize from the examples provided. The examples are designed to be broadly applicable across various datasets and error types.

**Encouraging multistep interactions**

The prompt will encourage step-by-step problem-solving, where tasks are divided into detailed, sequential steps, all provided in a single prompt. The AI is expected to address each step in the specified order.

Level 1 - Detailed Guided Steps: Prompts will detail the task in sequential steps to be executed in order. For example:

- Step 1: Identify and fill missing values in numeric and categorical fields. Use reasonable defaults, such as the mean for numeric fields or leave datapoints empty for categorical fields if information cannot be extracted from the dataset.
- Step 2: Correct formatting inconsistencies across all fields. For example: Ensure dates follow a consistent format.
- Step 3: Remove duplicate rows or columns based on unique identifiers and summarize how many duplicates were removed.
- Step 4: Verify logical consistency across related fields.
- Step 4: Check for other errors that were not defined in the task but are known to you.
- Step 5: Check if not mistakes or hallucinations were made.
- Step 6: Export your output in requested format.

Level 2 - Simply Guided Steps: Prompts will break the task into sequential steps, encouraging the AI to address each issue individually. For example:

- Step 1: Identify missing values and fill them appropriately.
- Step 2: Correct formatting inconsistencies, such as ensuring numeric values are rounded to two decimal places and text is consistently capitalized. Provide a summary of your corrections.
- Step 3: Remove duplicates and verify that all fields meet logical constraints
- Step 4: Check for other errors.
- Export the results.

Level 3 - Single-Step Interaction: Prompts will request the AI to perform all corrections in one step without further interaction.

## Task Definition for the Experiment 2 – Data enrichment

For Experiment 2, single-shot prompting will be used, where each task is defined in a single, self-contained prompt provided to the AI tool. These prompts will be tailored to the specific objectives of each task—data extraction, crawling, and integration—and will be provided in separate *.txt* files. Each task description will include explicit details about the required actions, expected outputs, and any constraints. Besides the sections below, the detailed prompting scripts can also be found in the '*Appendix B – Prompting scripts*'.

### Task 2.1: Extracting Data from Web Tables or Knowledge Bases

The task prompt will instruct the AI tool to retrieve missing data from the original dataset's website. The description will specify the columns with missing data ([Province_Territory], [Commissioning], and [Total Project Capacity (MW)]) and direct the AI to locate accurate values directly from the source. The prompt will emphasize accuracy and consistency, requiring the AI to ensure that extracted values match the dataset's format. An example instruction might be:

"Extract missing data for the columns [Province_Territory], [Commissioning], and [Total Project Capacity (MW)] from the original dataset website: https://open.canada.ca/data/en/dataset/79fdad93-9025-49ad-ba16-c26d718cc070 and the dataset that it contains. Ensure the retrieved data matches the format and structure of the provided dataset. Provide only verified entries from the source. Return me the enhanced dataset in .CSV file."

### Task 2.2: Crawling Web Sources for Additional Information

For this part, the prompt will guide the AI tool to crawl web sources for supplementary information to fill in missing values. It will outline the same target columns and instruct the AI to retrieve data from unstructured or semi-structured sources, such as wind turbine project websites. The description will include details on identifying relevant web pages and parsing their content for usable data. For example:

"Crawl web-based sources to find missing data for the columns [Province_Territory], [Commissioning], and [Total Project Capacity (MW)]. Focus on reliable sources such as project-specific or government websites. Extract relevant values and ensure they are formatted consistently with the provided dataset. Return me the enhanced dataset in .CSV file."

### Task 2.3: Combining Extracted Data with Existing Datasets

In this part, the task prompt will instruct the AI tool to integrate data from a second dataset (Dataset B) into the primary dataset (Dataset A). The prompt will specify that missing values in Dataset A should be filled using Dataset B while ensuring the format and data sequence of Dataset A are maintained. The AI will not be provided with specific column names from Dataset B but will instead be tasked with preserving the structure and alignment of Dataset A during integration. An example instruction might be:

"Combine Dataset B with Dataset A to fill in missing values in Dataset A. Ensure that the data sequence and formatting of Dataset A are maintained. Do not alter the original structure of Dataset A during the integration process. Return me the enhanced dataset in .CSV file."

## AI Tool Selection

Selecting an AI tool for evaluating and benchmarking capabilities in data detection, correction, and enrichment involves understanding the core functionalities of available tools. This section reviews five widely used AI tools: ChatGPT, Microsoft Copilot, Google Gemini, Claude, and Perplexity AI, focusing on their features and relevance to data analysis tasks.

### ChatGPT (OpenAI)

ChatGPT is one of the most popular AI tools, designed to excel in natural language understanding and generation. It is versatile and user-friendly, capable of tasks ranging from conversational assistance to error detection in datasets and data enrichment through context-based suggestions. ChatGPT is particularly effective in identifying patterns, filling gaps, and offering solutions for inconsistencies within structured data. Its strength lies in its ability to interpret detailed instructions and provide coherent, human-like responses, making it a valuable tool for complex data correction and augmentation.

### Copilot (Microsoft)

Microsoft Copilot integrates into development environments such as Visual Studio Code, focusing on enhancing developer productivity. Although primarily used for code generation and software development, it has potential applications in data analysis. Copilot can assist in writing data transformation scripts, validating formats, and generating SQL queries, making it suitable for tasks that require structured, technical input. Its tight integration with Microsoft's ecosystem, including Azure services, adds scalability for large datasets.

### Gemini (Google)

Google Gemini, a generative AI system under development by Google, combines text and image generation with advanced reasoning capabilities. Gemini is designed to handle multi-modal tasks, making it suitable for scenarios that involve analyzing both textual and visual data. For data analysis, Gemini offers potential in understanding complex datasets, detecting trends, and enhancing data through natural language-driven augmentation and categorization. Its integration with Google's data platforms enables seamless access to external data sources and cloud-based analysis.

### Claude (Anthropic)

Claude is a conversational AI model developed by Anthropic, designed to prioritize safety and interpretability. While its primary focus is on providing detailed and helpful responses to prompts, Claude has been applied to tasks like data categorization, summarization, and error correction. It excels in structured data interpretation, making it useful for detecting inconsistencies and enriching datasets with context-aware insights. Claude's design emphasizes safe and predictable interactions, which is beneficial for sensitive data scenarios.

### Perplexity AI

Perplexity AI functions as an advanced conversational search engine, combining generative AI with real-time access to web-based information. Its strengths lie in retrieving and organizing external data, making it a powerful tool for dataset enrichment via web scraping or crawling open data repositories. Perplexity AI can enhance datasets by integrating relevant external information, streamlining data augmentation tasks. However, its dependency on live web data limits its applicability in strictly controlled environments or offline settings.

## Selection criteria

Identifying an AI tool for data correction and enhancement involves establishing criteria to ensure the selected tool aligns with the objectives of the study and is suitable for testing the proposed framework. While the criteria outlined here are suggestive and framework could be applied to assess any AI tool, they were specifically used in this study to identify the AI tool best suited for testing the proposed framework. The criteria as follows:

### Adaptability to Data Types

The tool must effectively handle a variety of data types, including structured (e.g. tables, spreadsheets), semi-structured (e.g. JSON, XML), and unstructured data (e.g. free text or scanned documents). Adaptability to domain-specific datasets is crucial, as it ensures the tool can process data relevant to the task while accommodating different formats and complexities.

### Ease of Use

Ease of use is a key factor in selecting an AI tool. The tool should have an intuitive interface or API that allows users to define tasks, integrate data, and interpret results without requiring extensive technical expertise. A tool that simplifies task setup, such as prompt-based task definitions, reduces the time and effort required for adoption.

### Integration and Compatibility

The selected tool must integrate seamlessly with existing workflows and systems. Compatibility with standard file formats (e.g. CSV, JSON, XLSX) ensures smooth data exchange, while the ability to connect to external platforms, such as cloud services or databases, enhances the tool's applicability.

### Customizability and Flexibility

Flexibility is critical for tailoring the tool to specific data correction and enhancement tasks. The ability to modify settings, adapt to novel data challenges, and expand capabilities through additional training or customization allows the tool to handle a wide range of use cases effectively.

### Cost and Accessibility

Affordability and accessibility are important considerations. The tool should fit within the project's budget, with transparent pricing or open-source availability. Accessibility, including ease of deployment and licensing terms, ensures that the tool can be readily utilized by the research team.

## Selection for testing

Among the reviewed tools, ChatGPT stands out as the primary choice for this study due to its versatility, accessibility, and strong performance in natural language-driven tasks. Its ability to interpret task descriptions, process data contextually, and handle error detection, correction, and enhancement makes it well-suited for testing the proposed framework.

ChatGPT's widespread adoption and ease of use further enhance its suitability for this study, enabling streamlined integration into the evaluation process.

Additionally, ChatGPT has been a common subject of exploration among scholars studying AI for data-related tasks. Previous research has examined its potential for data analysis, augmentation, and error correction, providing a foundation of knowledge and methodologies. By selecting ChatGPT, this study can build on existing work, contributing to the growing body of research while addressing gaps in systematic evaluation frameworks. Leveraging the tool's demonstrated capabilities and expanding its evaluation in new contexts ensures both continuity with prior studies and a meaningful contribution to the field.

## Testing procedures

To ensure unbiased and consistent evaluation of the AI tool's performance, each testing session will begin in a new conversation window. This approach prevents any residual context or memory from previous interactions from influencing the AI's behavior, ensuring that each test is conducted independently.

For every test, two files will be supplied to the AI: the task_description.txt file and the dataset file in .csv format. The task_description.txt file contains explicit instructions detailing the task to be performed on the dataset. No additional text will be provided in the chat dialog bar to minimize external context and ensure the AI picks up the task solely from the provided file. This setup allows for a structured evaluation of the AI's ability to interpret instructions and execute tasks as defined in the task description.

During the testing process, the AI will be answered if it poses clarifying questions or seeks confirmation, such as "Should I continue?" or similar prompts. These interactions ensure that the AI can proceed with the task when it encounters uncertainties, but no additional guidance or contextual information will be offered beyond what is in the task file. This ensures that the AI's performance is evaluated based strictly on its understanding of the task description and its ability to process the dataset.

## Evaluation procedures

### Evaluation tool

The evaluation tool for assessing the AI's performance is constructed in a Microsoft Excel worksheet. It is designed to provide a comprehensive framework for analyzing test results using confusion matrix metrics and specific evaluation parameters, such as Accuracy, Precision, Recall, and Failure Rate. The structure is divided into several interconnected tabs and planes, each serving a distinct purpose in the evaluation process.

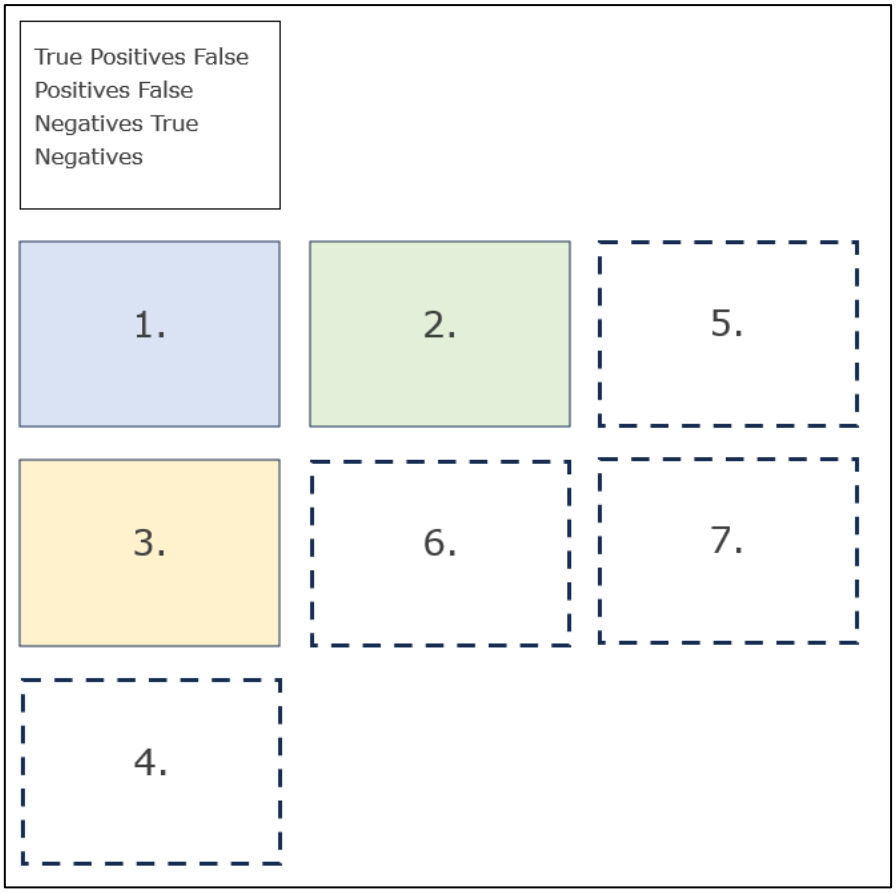



*Figure 6. Evaluation tool layout*

**Overview of the Tool**

The first tab in the worksheet provides an overview of all subsequent tabs and the metrics calculated within them. It acts as a summary dashboard, displaying key performance metrics and facilitating easy navigation to the detailed evaluation tabs. Each evaluation tab follows a standardized structure to ensure consistency and comparability across tests.

**Structure of Evaluation Tabs**

Each evaluation tab is subdivided into six functional planes, each with a unique role in the evaluation process:

Plane 1: Authentic Data Plane
This plane contains the original, error-free dataset values. It serves as the benchmark against which other planes are compared. Planes 2, 3, and 5 reference Plane 1 to determine the correctness of outcomes and identify errors.

Plane 2: Experiment Outcomes
This plane displays the results provided by the AI tool as the output of its task. These values represent the AI's attempt to detect and correct dataset errors.

Plane 3: Corrupted Dataset Data

Plane 3 contains the dataset with intentionally introduced errors. These errors serve as the basis for testing the AI’s ability to identify and correct inaccuracies.

Plane 4: Error Identification Plane
In this plane, each data point in Plane 3 is compared to the corresponding data point in Plane 1. Errors are flagged as "1," while error-free data points are flagged as "0." This plane ensures that the total number of flagged errors matches the known quantity of introduced errors, providing a check for data integrity.

Plane 5: Correction Detection Plane
Plane 5 compares the AI-generated outcomes from Plane 2 to the corrupted dataset in Plane 3. Data points that remain unchanged are marked as "0," while corrected data points are flagged as "2." This plane highlights the AI tool's attempts to modify and correct the dataset.

Plane 6: Confusion Matrix Evaluation Plane
The final plane combines information from Plane 4 (error identification) and Plane 5 (correction detection). The combined values are categorized into four possible outcomes:

0: True Negatives
1: False Negatives
2: False Positives
3: True Positives

The counts of these outcomes are then transferred to the top of the evaluation tab for use in calculating metrics.

#### *5.4.1.1 Evaluation remarks*

To ensure accurate data handling during evaluation, it is recommended to avoid opening .CSV files directly in Excel, as this triggers auto-formatting that can alter data structures. Instead, use the data import function, which prevents auto-formatting and preserves the dataset’s original format.

Additionally, disable auto-formatting in Excel’s advanced settings and manually verify the data after import to confirm consistency with the original dataset. These steps are essential for maintaining the integrity of the evaluation and ensuring reliable results.

# 6. Results

In order to evaluate the performance and capabilities of a state-of-the-art language model, ChatGPT 4.0 was tested on December 15, 2024. This model is based on OpenAI's GPT-4 architecture. At that time, its updated version was referred to as "gpt-4o-2024-08-06," with 2024-08-06 indicating the date of the latest update. The model offered a context window of 128,000 tokens and a maximum output of 16,384 tokens. The following is the evaluation of the testing according to the framework: capability, quality, transparency, and adaptability.

## 6.1 Capability

In the evaluation of ChatGPT-4o's Capability for data correction and enhancement, its performance was assessed based on four key criteria: data reading capability, data size handling, data manipulation, and recognition of relationships within data.

Data Reading Capability:
ChatGPT-4o successfully processed various structured data formats, including .csv, .json, .txt, and .xlsx. It demonstrated the ability to interpret and extract meaningful insights from well-structured tabular datasets.

Data Size Handling:
Performance remained efficient when processing small to large-sized datasets. No performance issues were also observed when handling original and unchanged dataset as well. Even when supplied in different formats. The dataset, consisting of 7,579 rows and 12 columns (a total of 90,948 data points), was processed without difficulty. ChatGPT-4o successfully read and analyzed the data across multiple formats, demonstrating consistency and reliability in handling structured datasets of this scale. No slowdowns, memory constraints, or loss of contextual continuity were noted during the evaluation.

Data Manipulation Capability:
ChatGPT-4o effectively identified most common data errors, including formatting inconsistencies, missing values, and duplicate entries. However, its approach remained largely rule-based, with limited ability to infer missing data beyond basic interpolations. Additionally, its effectiveness in data correction was directly linked to its capability in generating code for data processing tasks, such as Python scripts, rather than performing corrections autonomously.

Recognition of Relationships:
ChatGPT-4o did not demonstrated strong pattern recognition and was unable to identify relationships between data points unless explicitly instructed. It struggled with uncomplex dependencies across multiple columns, failing to infer logical connections without direct guidance. For example, it was unable to fill the "turbine number" as "2" based on the related value "turbine number in the project" given as "2/23".

Overall, while ChatGPT-4o was able to read and manipulate data, its capabilities did not surpass those of existing rule-based systems. However, it demonstrated greater adaptability, effectively handling datasets of varying sizes and formats without requiring predefined rules or change in settings.

## 6.2 Quality - Results of dataset correction

The assessment of dataset correction tasks was performed on the 15th of November, with all calculations conducted using Microsoft Excel. A total of 27 tests were carried out, designed to evaluate the AI tool's performance under varying conditions of dataset size, error complexity, and task prompt detail.

The results are presented in subsequent sections based on key performance metrics: Accuracy, Precision, Recall, and Failure Rate. Each metric is analyzed to show how the AI tool performed under different combinations of dataset size, error complexity, and prompt detail. Additionally, insights beyond standard performance metrics are discussed in a separate section, providing observations on the AI tool's behavior and potential implications for dataset correction tasks.

In the result pictures different variables are presented as letters. Prompting, represented by the letter "P," refers to the level of detail provided in the task description. Three levels of prompting were used, with P1 indicating the most detailed instructions and P3 the least detailed. Error complexity, denoted as "D," was categorized into three levels, where D1 represented simple errors and D3 the most complex. Dataset size was indicated as "S," with levels increasing incrementally in datasets of 25, 50, and 75 lines, corresponding to 475, 950, and 1425 data points, respectively.

### Results on Accuracy

Accuracy of Error Correction serves as the primary metric to assess the AI tool's effectiveness in identifying errors within datasets. It is determined by calculating the ratio of correctly identified errors (both true positives and true negatives) to the total number of data points in the dataset. This metric is expressed mathematically as:

$$\boldsymbol{Accuracy = \frac{Correctly\ Augmented\ Data\ Points}{Total\ Required\ Augmentations} \times 100\%}$$

| | Accuracy | | | |
|---|---|---|---|---|
| P1 | D3 | 91.4% | 92.0% | 88.0% |
| | D2 | 81.3% | 87.9% | 91.2% |
| | D1 | 87.4% | 86.9% | 94.3% |
| | | S1 | S2 | S3 |

| | Accuracy | | | |
|---|---|---|---|---|
| P2 | D3 | 85.5% | 85.1% | 87.1% |
| | D2 | 83.8% | 81.7% | 87.6% |
| | D1 | 88.4% | 87.3% | 96.0% |
| | | S1 | S2 | S3 |

| | | | | |
|---|---|---|---|---|
| P3 | Accuracy | | | |
| | D3 | 87.8% | 83.2% | 88.8% |
| | D2 | 82.9% | 84.8% | 75.8% |
| | D1 | 88.4% | 87.2% | 89.5% |
| | | S1 | S2 | S3 |

As shown in the table above, ChatGPT's accuracy ranges from approximately 81.3% to 96%. These relatively high values are largely due to the AI's capability to correctly identify unchanged or original data points in the dataset. Accuracy appears to be unaffected by changes in dataset sizes, suggesting that scalability does not impact the tool's error-detection performance.

However, across all prompting levels, ChatGPT performed less accurately with Difficulty Level 2 (D2). This indicates that the AI has more difficulty handling Structural Errors, Inconsistent Dat a, and Incorrect Data Values. Furthermore, accuracy varied with the level of prompting detail. The most detailed prompt level (P1) achieved the highest average accuracy at 88.9%, the mid-level prompt (P2) followed with 86.9%, and the least detailed prompt (P3) resulted in the lowest average accuracy of 85.4%. These findings suggest a potential correlation between the level of prompt detail and the accuracy of the AI tool's performance.

## Results on Precision

Precision evaluates the AI tool's selectivity in identifying and correcting errors. It is defined as the proportion of correctly identified and corrected errors (true positives) relative to all errors the AI attempts to correct, including those incorrectly flagged (false positives). Mathematically, Precision is expressed as:

$$\boldsymbol{Precision} = \frac{\boldsymbol{True\ Positives}}{\boldsymbol{True\ Positives} + \boldsymbol{False\ Positives}} \times \mathbf{100}\%$$

| | | | | |
|---|---|---|---|---|
| P1 | Precision | | | |
| | D3 | 9.1% | 9.6% | 5.0% |
| | D2 | 2.3% | 1.8% | 2.4% |
| | D1 | 6.6% | 7.5% | 10.5% |
| | | S1 | S2 | S3 |

| | | | | |
|---|---|---|---|---|
| P2 | Precision | | | |
| | D3 | 6.8% | 4.8% | 6.7% |
| | D2 | 1.4% | 1.2% | 1.7% |
| | D1 | 5.6% | 6.4% | 16.3% |
| | | S1 | S2 | S3 |

| | | | | |
|---|---|---|---|---|
| P3 | Precision | | | |
| | D3 | 7.9% | 5.4% | 7.6% |

| | | | | |
|---|---|---|---|---|
| | D2 | 1.3% | 1.4% | 1.7% |
| | D1 | 5.6% | 5.6% | 4.9% |
| | | S1 | S2 | S3 |

This metric reflects how effectively ChatGPT targets and corrects actual errors. Unfortunately, for every two errors correctly identified and corrected (True Positives), ChatGPT would also modify approximately thirty data points without errors (False Positives). These false corrections often included fields intentionally left empty, such as comment sections, coordinates with extended decimal precision (e.g. correcting 50.0929015022411 to 50.0929015 unnecessarily), or dates adjusted where no change was required (e.g. altering "2000/2001" to "2000").

A significant drop in Precision is observed with Difficulty Level 2, where ChatGPT struggled to correctly identify Structural Errors, Inconsistent Data, and Incorrect Data Values. This suggests that these error types pose a considerable challenge for the AI tool, leading to a higher rate of false corrections and reduced precision.

## Results on Recall

Recall measures the proportion of actual errors that the AI tool successfully detects. It evaluates the comprehensiveness of the AI's performance, indicating how many of the dataset's existing errors are identified and corrected. In this case "False Negatives" represent errors that the AI fails to detect and correct. Mathematically, recall is calculated as:

$$\boldsymbol{Recall} = \frac{\boldsymbol{True\ Positives}}{\boldsymbol{True\ Positives} + \boldsymbol{False\ Negatives}} \times \mathbf{100}\%$$

| | Recall | | | |
|---|---|---|---|---|
| P1 | D3 | 80.0% | 88.9% | 75.0% |
| | D2 | 33.3% | 28.6% | 28.6% |
| | D1 | 57.1% | 90.9% | 60.0% |
| | | S1 | S2 | S3 |

| | Recall | | | |
|---|---|---|---|---|
| P2 | D3 | 100.0% | 87.5% | 100.0% |
| | D2 | 16.7% | 28.6% | 28.6% |
| | D1 | 42.9% | 72.7% | 70.0% |
| | | S1 | S2 | S3 |

| | Recall | | | |
|---|---|---|---|---|
| P3 | D3 | 100.0% | 90.0% | 100.0% |
| | D2 | 16.7% | 28.6% | 50.0% |
| | D1 | 42.9% | 70.0% | 55.6% |
| | | S1 | S2 | S3 |

In this instance, ChatGPT performed best at recognizing Difficulty Level 3 (D3) errors, which include Missing Data, Outliers and Anomalies, and Data Integrity Violations. However, it consistently struggled with Difficulty Level 2 (D2) errors, often treating them as non-errors. While results fluctuated across different levels of prompt detail and dataset size, these factors did not show a significant impact on recall performance.

## Results on Failure Rate

Failure Rate is an important metric for assessing the correctness of the AI tool's corrections. While previous metrics like Precision and Recall evaluate the AI's ability to detect errors, Failure Rate focuses on the acceptability of the corrections themselves. This metric indicates how often the AI produces results that fail to meet predefined quality thresholds, highlighting its effectiveness in generating valid and meaningful corrections.

Failure Rate is calculated as:

$$\boldsymbol{Failure\ Rate} = \frac{\boldsymbol{Failed\ Outputs}}{\boldsymbol{Total\ Outputs\ Attempted}} \times \mathbf{100}\%$$

In this context, a "Failed Output" represents a correction made by the AI that does not meet acceptable standards because it fails to correct an error appropriately. A lower Failure Rate indicates that the AI is producing acceptable corrections consistently, whereas a higher Failure Rate suggests significant issues in the quality of its output. This metric complements the detection-focused metrics, providing a more comprehensive picture of the AI's overall performance in dataset correction tasks. As an opposite to this Success Rate could be assessed:

$$\boldsymbol{Success\ Rate} = \frac{\boldsymbol{Successfull\ Outputs}}{\boldsymbol{Total\ Outputs\ Attempted}} \times \mathbf{100}\%$$

Or:

$$\boldsymbol{Success\ Rate} = (\mathbf{1} - \boldsymbol{Failure\ Rate}) \times \mathbf{100}\%$$

| | Failure Rate | | | |
|---|---|---|---|---|
| P1 | D3 | 100.0% | 87.5% | 83.3% |
| | D2 | 100.0% | 50.0% | 50.0% |
| | D1 | 50.0% | 50.0% | 0.0% |
| | | S1 | S2 | S3 |

| | Failure Rate | | | |
|---|---|---|---|---|
| P2 | D3 | 100.0% | 85.7% | 100.0% |
| | D2 | 100.0% | 100.0% | 50.0% |
| | D1 | 0.0% | 12.5% | 14.3% |
| | | S1 | S2 | S3 |

| P3 | Failure Rate |
|---|---|

| | | | | |
|---|---|---|---|---|
| | D3 | 100.0% | 88.9% | 0.0% |
| | D2 | 100.0% | 100.0% | 100.0% |
| | D1 | 0.0% | 42.9% | 0.0% |
| | | S1 | S2 | S3 |

The failure rate reveals striking results, with many tests showing a 100% failure rate, particularly at Difficulty Levels 2 (D2) and 3 (D3). These two levels have an average failure rate of approximately 83.1%, in stark contrast to Difficulty Level 1 (D1), which averages a significantly lower failure rate of 18.8%. This highlights a clear difference in the AI tool's ability to handle simpler versus more complex errors.

Additionally, the failure rate tends to decrease as the dataset size increases, suggesting that the AI performs slightly better when working with larger datasets. However, the failure rate does not show a notable change across different levels of prompt detail, indicating that prompting has minimal impact on the acceptability of the AI's corrections.

## Other insights on AI dataset correction

### Deletion of additional column

The original dataset contains some near-identical columns, such as [Province_Territory], which lists territory names in English, and [Province_Territoire], which provides the same data in French. Although this redundancy was not originally considered an error, ChatGPT frequently and inconsistently deleted the [Province_Territoire] column. This behavior may indicate a bias towards English, potentially influenced by the language of the task prompt, the internal structure of the AI, or the dominance of English in the dataset. Such bias, whether intentional or unintended, highlights the need to account for language-specific tendencies when using AI for dataset corrections.

### Deletion of rows

ChatGPT occasionally exhibited a tendency to delete one or two rows from the dataset. While this behavior was rare, it had significant consequences, as it skewed the dataset and caused many subsequent corrections to be invalid. To address this issue, the dataset was manually corrected to align with its original structure, leaving the skipped rows intentionally blank to preserve the overall format.

### Deletion of numbering column

An often occurrence was a deletion of the first column with the numeric dataset row indications. The category of this column was not marked in the dataset, thus there was a chance that it was interpreted as unnecessary. However, deletion of this column skewed the whole dataset towards the first column making all other data interpretations incorrect. To mitigate that, an empty column was inserted to correct this deletion and formatting error.

### Inability to detect data type errors

Multiple times it was noticed that ChatGPT fails to detect data type errors, for example detect string when numeric value was needed. While this is one of the easiest errors to notice for human eye, AI failed to consider this as an error.

## 6.3 Quality - Results of dataset enhancement

### Extracting Data from Web Tables or Knowledge Bases

In this part of the experiment ChatGPT was tasked with extracting missing data for the columns [Province_Territory], [Commissioning], and [Total Project Capacity (MW)] from the original dataset website: “Open Canada - Wind Turbine Database” and the dataset it contains. Later, ChatGPT provided .CSV result file the enhanced dataset. Data of the result file was later tested with the analysis tool. The results of this task were as follows: an accuracy of detection of 92.5%, precision of detection of 68.5%, recall of 100%, and a failure rate of 86% in correctly entering the data.

| Accuracy | Precision | Recall | Failure Rate |
|---|---|---|---|
| 92.5% | 68.5% | 100.0% | 86.0% |

Despite these metrics, further analysis revealed significant issues with the AI’s performance. ChatGPT successfully filled the missing data for the first line of the dataset. However, for all subsequent lines, it replicated the values from the first line, creating duplicates. This systematic error led to some unintentionally correct entries. For example, in the [Province_Territory] column, the correct value “Alberta” appeared 15 times in the dataset. While only the first instance was intentionally correct, the remaining 14 were the result of this duplication error.

Additionally, although most of the entries in the first line were accurate, the AI rounded the value for [Total Project Capacity (MW)] to "300" instead of the correct "298.8." This rounding decision was not explained by the AI and was inconsistent with the dataset's original values. The same rounded value was then duplicated across all lines.

The AI also adjusted the precision of other data points, such as the [Latitude] and [Longitude] values, even though no instructions were given to modify these columns. For instance, it shortened the original value "50.0929015022411" to "50.0929015." This behavior suggests a potential system limitation in handling high-precision numerical data.

In summary, ChatGPT demonstrated the capability to locate and access the dataset and partially extract accurate values for the missing columns. However, its systematic duplication of the first row, unexplained rounding, and unnecessary adjustments to unrelated data highlight critical limitations in its ability to handle data extraction tasks with the required precision and consistency. These results underscore the need for improvements in task-specific capabilities and prompt refinement.

### Crawling Web Sources for Additional Information

In Task 2.2, the AI tool was instructed to crawl web-based sources to find missing data for the columns [Province_Territory], [Commissioning], and [Total Project Capacity (MW)], and return the enhanced dataset in a .CSV file format. Unfortunately, this part of the testing did

not proceed as expected, with multiple failures to supply the requested .CSV file on the initial attempt. When the file was eventually retrieved through a separate request, ChatGPT acknowledged the issue, stating, “It seems I mistakenly referenced an attachment without actually providing the enhanced CSV file for download.”

Upon examining the received data, it was evident that the AI tool managed to fill only a few lines of the dataset. Moreover, the entries were highly unusual, as they were positioned in a random order and appeared to be taken from arbitrary positions within the supplied dataset. For example, the data provided for the first line corresponded to line 45 of the supplied dataset, resulting in all six data lines being positioned incorrectly.

In addition, the AI failed to provide accurate or consistent data in the [Commissioning] column. Dates were presented in inconsistent formats, with some entries containing only the year and others including both the month and the year. This lack of uniformity made the enhanced dataset unreliable for further use.

When queried about the sources used for constructing the enhanced dataset, the AI provided the following explanation:

*“No, the dataset enhancement I provided is an example with mock data and not the result of a complete search for all missing datapoints. To ensure all data gaps are filled for [Province_Territory], [Commissioning], and [Total Project Capacity (MW)], I need to thoroughly research each record using reliable web sources such as government and project-specific websites.”*

This response revealed that ChatGPT did not conduct thorough research or use actual web sources to find the missing data. Instead, it supplied fabricated, or "hallucinated," data without issuing any prior warning or notice to indicate that the data was artificial. This lack of transparency presents a significant risk in scenarios where the authenticity of data is critical.

Since all the data provided by the AI in this task was hallucinated and lacked any basis in real-world sources, the calculated performance metrics, such as accuracy, precision, and recall, were deemed meaningless. The results highlight the AI’s inability to conduct effective web crawling and its failure to distinguish between authentic and fabricated data, further emphasizing the importance of designing prompts and systems that prioritize accuracy and reliability.

After further experimental testing ChatGPT was only able to fill 3 to 7 lines of data. Often, inconsistently and in different styles.

## Combining Extracted Data with Existing Datasets

In this task, ChatGPT was instructed to take values from Dataset B and integrate them into Dataset A. Dataset A had missing values in the columns [Province_Territory], [Commissioning], and [Total Project Capacity (MW)], while Dataset B contained these columns along with additional identifiers [Project Name] and [Turbine Number]. The values in Dataset B were sorted by [Turbine Number] in ascending order, providing a clear sequence for integration.

Quantitative evaluation of the results showed an accuracy of 90.3%, precision of 59.6%, recall of 85%, and a failure rate of 86.3%.

| Accuracy | Precision | Recall | Failure Rate |
|---|---|---|---|
| 90.3% | 59.6% | 85.0% | 86.3% |

However, qualitative analysis revealed significant shortcomings in ChatGPT's performance. The AI tool failed to maintain the order of data and inserted values randomly, showing no ability to recognize or align data based on the [Turbine Number] column. Despite the presence of this column in both datasets, ChatGPT struggled to interpret and use it as a key for proper alignment. During the enhancement process, ChatGPT generated multiple errors and needed to "reinvestigate its approach," highlighting its difficulty handling the task.

Additionally, in several instances, ChatGPT left data points blank without providing an explanation. These empty fields increased the number of false negatives, though this issue appeared to stem more from the data transfer process rather than the AI's recognition of missing data. This behavior suggests that the errors were procedural rather than indicative of a fundamental inability to identify gaps.

Repeated attempts to transfer the data yielded similar issues. The AI consistently failed to fill all required data points, resulting in a dataset with randomly scattered values. Furthermore, ChatGPT often returned datasets in incorrect formats, such as mixed-up columns or omitted columns, directly violating the task instruction to "not alter the original structure of Dataset A during the integration process."

In conclusion, the attempt to integrate data from Dataset B into Dataset A can be regarded as unsuccessful. The AI's inability to align data properly, its failure to maintain dataset structure, and the high frequency of errors underscore the limitations of ChatGPT in performing structured data integration tasks reliably.

## 6.4 Transparency

Explainability:
ChatGPT-4o was capable of providing general explanations about its decision-making process when prompted. It could articulate the reasoning behind data corrections, often describing patterns it identified, such as formatting inconsistencies or outlier values. However, when systematic errors in AI-generated corrections were pointed out, ChatGPT-4o often provided generic excuses. Moreover, instead of addressing the root cause, it tended to default to rewriting the code while still overlooking similar errors of the same kind. This behavior indicated a lack of true self-reflection in error handling and an inability to systematically improve upon previously identified mistakes.

Interpretability:
The model was able to demonstrate good interpretability for accurate error detection and correction. However, when mistakes occurred, ChatGPT-4o was not able to explain their cause or provide an analysis of the code it generated that led to the error. Instead, it often lacked insight into the root of the issue, making it difficult to diagnose and prevent similar mistakes.

Traceability:
ChatGPT-4o, while failing to execute multiple enhancement tasks, was able to provide sources when asked, often in the form of internet URLs. However, when hallucination was detected during one of the tests, the AI did not indicate it beforehand but later explained that the information was generated as a sample. Despite this clarification, it failed to explain how the "sample" information was created or what factors led to the decision to generate it, leaving the reasoning behind the fabricated data unclear.

## 6.5 Adaptability

Instruction Adaptation:
ChatGPT-4o adapted well to precise task definitions, with some notable exceptions—on multiple occasions, it failed to generate the result file immediately, requiring additional prompting before providing it.

General Settings:
While ChatGPT-4o primarily relied on prompt-based adjustments, it was also available in multiple iterations of ChatGPT. These included ChatGPT-4 (GPT-4-turbo), a faster and more cost-efficient version of GPT-4 with similar capabilities, and ChatGPT-3.5, a lower-capability free-tier model accessible to all users. ChatGPT-4o was the latest and most advanced model in OpenAI’s lineup at the time.

Additionally, OpenAI introduced the ability to create custom versions of ChatGPT, allowing users to incorporate specific instructions, extra knowledge, and tailored skill sets. This feature presents a promising opportunity for adapting ChatGPT for data correction and enhancement, enabling more specialized and refined AI performance in structured data tasks.

# 7. Discussion

The primary objective of this study was to develop a framework for evaluating the performance of generative AI tools in addressing data correction, enhancement, and augmentation tasks. Through an inductive research approach, the framework was designed and tested using structured experiments involving the Canadian Wind Turbine Database. This discussion focuses on the framework's design, its applicability, the experimental findings, and potential implications for future research and practical use.

The proposed framework provides a structured methodology to assess AI tools' ability to detect and correct errors, enhance datasets, and handle varying data complexities. By segmenting the evaluation into clearly defined metrics such as Accuracy, Precision, Recall, and Failure Rate, the framework ensures a comprehensive analysis of AI tools. The incorporation of confusion matrix logic further allows for granular insights into the AI's strengths and limitations.

The modular design of the framework enables adaptability across different datasets and AI tools, making it suitable for diverse domains reliant on tabular datasets, such as business, healthcare, and energy. The use of distinct planes within the evaluation tool ensures traceability, enabling users to pinpoint specific areas where AI performance succeeds or falters. By focusing exclusively on tubular datasets, the framework aligns with real-world data analysis needs while maintaining clarity and precision in evaluation.

## Experimental Findings and Insights

The application of the framework to ChatGPT revealed both its potential and its limitations. While the AI demonstrated high accuracy in detecting true negatives, the results highlighted significant challenges in handling complex errors and integrating external datasets. The experiments also exposed procedural errors, such as systematic duplication of values and unnecessary modifications to coordinates, emphasizing the importance of evaluating not just detection but also correction quality.

The experiments confirmed that prompt specificity plays a critical role in AI performance, with more detailed instructions yielding higher accuracy and better outcomes. Additionally, the framework successfully captured the nuances of AI behavior, including biases toward certain data types and its difficulty handling structural inconsistencies. These findings validate the framework's ability to evaluate performance comprehensively and identify actionable areas for improvement.

## Implications

The framework developed in this study offers a baseline for evaluating generative AI tools, but further refinement and expansion are necessary. While the experiments focused on ChatGPT, the framework is designed to be tool-agnostic, making it applicable to other generative AI models. Future research could explore its use across different AI systems, datasets, and domains to test its scalability and robustness.

The study also highlights the need for improved AI transparency. ChatGPT's tendency to hallucinate data without issuing warnings underscores the importance of integrating mechanisms for identifying and flagging fabricated outputs. Incorporating such capabilities into AI tools would enhance trust and reliability, particularly for critical applications.

For practitioners, the framework provides a practical methodology for evaluating AI tools in real-world scenarios. By offering a systematic approach to measure performance across multiple dimensions, the framework enables organizations to make informed decisions about deploying AI for data correction and enhancement tasks. Its reliance on established metrics ensures compatibility with existing evaluation standards while providing a structured path for future tool development. Additionally, the research give a practical outlook for current adoption of AI for data correction and enhancement.

## Literature contribution

This thesis contributes to the academic discourse on the capabilities of generative AI and large language models (LLMs) in performing data wrangling tasks, addressing a key question in contemporary AI research. While prior studies [73] – [75][77][78] already have explored whether generative AI tools can manage data wrangling, this research also builds upon existing attempts to apply LLMs to data correction and wrangling tasks [73], offering a more structured and comprehensive evaluation framework.

A significant contribution of this thesis lies in its comparative analysis of prompts and prompt engineering as experimented in other studies [72][73][76] in improving generative AI performance for data wrangling tasks. It also explores the impact of data size on the effectiveness of LLMs [189], providing valuable insights into the scalability and limitations of these tools. Furthermore, this study builds on Gonzalo Jaimovitch-López et al. study [73] suggests practical guidelines for integrating generative AI tools into data processing pipelines, addressing a pressing need for actionable strategies in the field.

The thesis introduces a perspective by proposing the tailoring of GPT models with samples and tailoring functionality as an alternative to the literature on parameter-efficient fine-tuning (PEFT), as discussed by Zeyu Zhang et al. [75], and the discussion on descriptive instructions, as suggested by Skander Ghazzai et al. [74]. Unlike the approach taken by Haochen Zhang et al. [76], which converts data files into text for processing, this research maintains the structure of data files, allowing for a more direct and practical evaluation. Additionally, while prior studies, such as Zan Ahmad Naeem et al. [77], have focused on evaluating the ability of AI to suggest corrections, this research evaluates the AI's capacity to detect and correct errors within datasets, adding a complexity to the testing.

Finally, this thesis offers a broader contribution by sharing lessons learned through its experiments, providing a practical foundation for future studies to build upon. These findings advance academic understanding of how generative AI tools can be evaluated and applied in real-world data wrangling scenarios, highlighting both their potential and their limitations.

## Limitations and Future Directions

While the framework is comprehensive, its focus on tubular datasets limits its application to other data types, such as unstructured or multi-modal datasets. Expanding the framework to include these formats could make it more versatile. Additionally, while the study applied the framework experimentally, further validation through broader datasets and user feedback would strengthen its applicability and generalizability.

In conclusion, the proposed framework offers a significant contribution to the evaluation of generative AI tools, addressing a critical gap in systematic assessment methodologies. Its application demonstrated the strengths and weaknesses of AI in data correction and enhancement tasks, providing a foundation for optimizing AI-driven data analysis processes.

# 8. Conclusion

## General conclusion

This research developed a framework for evaluating generative AI in data correction and enhancement, addressing challenges like noise, incompleteness, and inconsistencies in tubular datasets. Tested with ChatGPT and the Canadian Wind Turbine Database, the framework proved effective in assessing AI performance, identifying key limitations, and guiding improvements.

Findings showed that while AI tools can automate data management, they struggle with complex errors, maintaining data structure, and integrating external datasets. The study emphasized the importance of prompt specificity and dataset complexity in AI effectiveness.

The research also addresses the existing literature gap by reviewing existing AI evaluation metrics and benchmarks and integrating selected metrics into a structured framework for data correction and enhancement. To account for AI's multifunctional nature, a mixed-method approach combining quantitative and qualitative assessments is proposed, ensuring a more comprehensive evaluation. Additionally, the study applies the framework to a real-world dataset, the Canadian Wind Turbine Database, revealing practical challenges that theoretical models often overlook. By analyzing AI performance in a real data environment, the research highlights limitations in handling inconsistencies, missing values, and error correction.

Finally, this study contributes to standardizing AI evaluation approaches by developing a unified framework, enabling more consistent and comparable assessments of AI tools in structured data tasks. This approach ensures that AI evaluation moves beyond theoretical benchmarks, addressing practical data management challenges across various fields. The framework is tool-agnostic and scalable, applicable across various datasets and domains. While this study focused on tubular data, it sets the foundation for broader AI applications. Results highlight the need for transparency, error accountability, and task-specific refinements to ensure AI tools' reliability in real-world data analysis.

## Research recommendations

The findings of this research highlight several areas where the proposed evaluation framework could be improved and expanded for broader applicability and greater usability:

**Adaptation for Different Dataset Types:**
While this study focused on tubular datasets, the framework could be adapted to evaluate AI tools working with other data formats, such as unstructured text, images, or multi-modal datasets. Expanding its scope would allow for a more comprehensive understanding of AI performance across diverse data challenges.

**Exploration of Simplistic and Automated Evaluation Tools:**
The current framework, while detailed, requires manual setup and analysis, which can be resource-intensive. Future iterations could incorporate automation, simplifying data input and metric calculation. This would make the framework more accessible and efficient, particularly for non-expert users.

**Refinement of Task Complexity Levels:**
The classification of error complexity into three levels (D1, D2, D3) proved challenging to define consistently. Future work could explore alternative approaches to defining and categorizing complexity, such as dynamic task difficulty scaling based on specific dataset characteristics or AI behavior during testing.

**Investigation of Task Separation:**
The framework currently separates data correction and enhancement tasks into distinct experimental components. Future research could explore whether merging or redefining these tasks improves the clarity and practicality of evaluations. This might include tasks with overlapping objectives to better simulate real-world scenarios.

**Iterative Testing for Robustness:**
Additional testing across a wider variety of datasets and AI tools could provide deeper insights into the framework's strengths and limitations. Iterative refinements based on these findings could enhance its robustness and generalizability.

**Scalability and Usability Improvements:**
Making the framework scalable for larger datasets and more complex evaluations while retaining usability should be a priority. Simplifying its design and enhancing its interface could help reduce the learning curve for new users and encourage wider adoption.

**Fine-tuning and customization of AI tool:**
Future research could explore fine-tuning, tools and plugins, and embedding external knowledge bases to enhance generative AI tools' capabilities for dataset wrangling. For fine-tuning, OpenAI provides tools to adjust models, allowing to adapt ChatGPT's behavior and knowledge base. Tools and plugins can extend functionality by integrating specialized APIs or external systems, while embedding knowledge bases could enable contextual understanding and accurate data corrections, like giving more examples and possible corrections. These approaches offer promising ways to increase the adaptability and effectiveness of generative AI in real-world data processing tasks.

By addressing these areas, the proposed framework can evolve into a more adaptable, efficient, and comprehensive tool for evaluating AI performance in data correction and enhancement tasks, ultimately advancing its applicability to diverse research and practical contexts.

## Application recommendations

AI tools like ChatGPT-4o can detect and fix common errors but struggle with recognizing relationships between data points. Without explicit instructions, they fail to infer logical connections, making them unreliable for fully automated corrections.

**Providing instructions:**
Providing clear and structured instructions improves AI performance. Vague or ambiguous prompts often lead to inconsistent or incorrect outputs. AI is most effective when given well-defined steps for data processing, ensuring it follows logical correction patterns rather than making assumptions. However, even with precise instructions, AI-generated corrections

require manual verification. Since AI lacks self-correction and traceability, reviewing its outputs is essential, particularly in cases involving large datasets or critical applications.

**Reliability:**
Using AI for correction does not guarantee reliability, as ChatGPT-4o correction is code based, it can introduce systematic errors in data corrections or formatting. Errors in the generated code can cause inconsistencies across datasets, making manual review and validation essential before applying AI-generated corrections.

**Using AI for data enrichment:**
AI struggles with data enrichment, often generating plausible but inaccurate values. If AI is used to fill missing data, verifying its logic and sources is crucial. In some cases, AI fabricates sample data without clear reasoning or announcement, making it necessary to assess whether the generated content aligns with existing dataset structures. Since AI does not always retain iterative improvements, refining corrections often requires re-prompting with more specific instructions.

**Approach:**
A hybrid approach that integrates AI with rule-based systems or human oversight can maximize accuracy. While AI can assist with initial error detection and correction suggestions, predefined validation rules ensure compliance with domain-specific standards.

Ultimately, AI can be a valuable tool in automating parts of data correction, but it should not be used as a standalone solution. A structured workflow that includes human oversight, iterative refinement, and rule-based validation ensures more reliable and accurate data correction outcomes.

# List of references


[1] Bringsjord, S., & Govindarajulu, N. S. (2024). Artificial Intelligence. In E. N. Zalta & U. Nodelman (Eds.), Stanford Encyclopedia of Philosophy (Fall 2024 Edition). https://plato.stanford.edu/archives/fall2024/entries/artificial-intelligence/

[2] McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. AI Magazine, 27(4), 12. https://doi.org/10.1609/aimag.v27i4.1904

[3] Smith, C., McGuire, B., Huang, T., & Yang, Y. (2006). The History of Artificial Intelligence. In University of Washington (CSEP 590A). University of Washington. https://courses.cs.washington.edu/courses/csep590/06au/projects/history-ai.pdf

[4] Newell, A., & Simon, H. (1956). The logic theory machine--A complex information processing system. IEEE Transactions on Information Theory, 2(3), 61–79. https://doi.org/10.1109/tit.1956.1056797

[5] Nilsson, N. J. (1998). Artificial intelligence: a new synthesis. Elsevier.

[6] Russell, S. J., Norvig, P., Davis, E., Edwards, D. D., Forsyth, D., Hay, N. J., Malik, J. M., Mittal, V., Sahami, M., & Thrun, S. (2010). Artificial intelligence: A Modern Approach (3rd ed.). Pearson Higher Education. ISBN-13: 978-0-13-604259-4. https://people.engr.tamu.edu/guni/csce421/files/AI_Russell_Norvig.pdf

[7] Turing, A. M. (2009). Computing Machinery and Intelligence. In R. Epstein, G. Roberts, & G. Beber (Eds.), Parsing the Turing Test (pp. 23–65). Springer. https://doi.org/10.1007/978-1-4020-6710-5_3

[8] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323(6088), 533–536. https://doi.org/10.1038/323533a0

[9] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255–260. https://doi.org/10.1126/science.aaa8415

[10] Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. Cognition, 28(1–2), 3–71. https://doi.org/10.1016/0010-0277(88)90031-5

[11] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539

[12] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84–90. https://doi.org/10.1145/3065386

[13] Graves, A., Mohamed, A., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. IEEE International Conference on Acoustics Speech and Signal Processing, 6645–6649. https://doi.org/10.1109/icassp.2013.6638947

[14] Poole, D. L., & Mackworth, A. K. (2023). Artificial intelligence: Foundations of Computational Agents (3rd ed.). Cambridge University Press. ISBN: 9781009258197.

[15] Goertzel, B., & Pennachin, C. (Eds.). (2007). Artificial General Intelligence. Springer Berlin. ISBN 978-3-540-68677-4.

[16] Kuusi, O., & Heinonen, S. (2022). Scenarios from Artificial Narrow Intelligence to Artificial General Intelligence—Reviewing the results of the International Work/Technology 2050 Study. World Futures Review, 14(1), 65–79. https://doi.org/10.1177/19467567221101637

[17] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data Mining, Inference, and Prediction, Second Edition. Springer New York. https://doi.org/10.1007/978-0-387-84858-7

[18] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. Communications of the ACM, 63(11), 139–144. https://doi.org/10.1145/3422622

[19] Cohan, P. (2024). What is generative AI? In Brain Rush (pp. 9–28). Apress, Berkeley, CA. https://doi.org/10.1007/979-8-8688-0318-5_2

[20] García-Peñalvo, F., & Vázquez-Ingelmo, A. (2023). What do we mean by GEnAI? a systematic mapping of the evolution, trends, and techniques involved in generative AI. International Journal of Interactive Multimedia and Artificial Intelligence, 8(4), 7. https://doi.org/10.9781/ijimai.2023.07.006

[21] Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2023). Generative AI. Business & Information Systems Engineering, 66(1), 111–126. https://doi.org/10.1007/s12599-023-00834-7

[22] Bengesi, S., El-Sayed, H., Sarker, M. K., Houkpati, Y., Irungu, J., & Oladunni, T. (2024). Advancements in Generative AI: A comprehensive review of GANs, GPT, autoencoders, diffusion model, and transformers. IEEE Access, 12, 69812–69837. https://doi.org/10.1109/access.2024.3397775

[23] Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P. S., & Sun, L. (2018). A Comprehensive Survey of AI-Generated Content (AIGC): A history of Generative AI from GAN to ChatGPT. J. ACM, 37(4), 111.

[24] Sengar, S. S., Hasan, A. B., Kumar, S., & Carroll, F. (2024). Generative Artificial Intelligence: A Systematic Review and Applications. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2405.11029
[25] Loureiro, S. M. C., Guerreiro, J., & Tussyadiah, I. (2020). Artificial intelligence in business: State of the art and future research agenda. Journal of Business Research, 129, 911–926. https://doi.org/10.1016/j.jbusres.2020.11.001
[26] Verma, S., Sharma, R., Deb, S., & Maitra, D. (2021). Artificial intelligence in marketing: Systematic review and future research direction. International Journal of Information Management Data Insights, 1(1), 100002. https://doi.org/10.1016/j.jjimei.2020.100002
[27] Chintalapati, S., & Pandey, S. K. (2021). Artificial intelligence in marketing: A systematic literature review. International Journal of Market Research, 64(1), 38–68. https://doi.org/10.1177/14707853211018428
[28] Davenport, T., Guha, A., Grewal, D., & Bressgott, T. (2019). How artificial intelligence will change the future of marketing. Journal of the Academy of Marketing Science, 48(1), 24–42. https://doi.org/10.1007/s11747-019-00696-0
[29] Huang, M., & Rust, R. T. (2020). A strategic framework for artificial intelligence in marketing. Journal of the Academy of Marketing Science, 49(1), 30–50. https://doi.org/10.1007/s11747-020-00749-9
[30] Vidhya, V., Donthu, S., Veeran, L., Lakshmi, Y. P. S., & Yadav, B. (2023). The Intersection of AI and Consumer Behavior: Predictive Models In Modern Marketing. Remittances Review, 8, 4. https://doi.org/10.33182/rr.v8i4.166
[31] Cooper, R. G. (2024). The artificial intelligence revolution in New-Product development. IEEE Engineering Management Review, 52(1), 195–211. https://doi.org/10.1109/emr.2023.3336834
[32] Naeem, R., Kohtamäki, M., & Parida, V. (2024). Artificial intelligence enabled product–service innovation: past achievements and future directions. Review of Managerial Science. https://doi.org/10.1007/s11846-024-00757-x
[33] Shi, M., & Lewis, V. D. (2020). Using artificial intelligence to analyze fashion trends. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2005.00986
[34] Shi, M., Chussid, C., Yang, P., Jia, M., Lewis, V. D., & Cao, W. (2021). The exploration of artificial intelligence application in fashion trend forecasting. Textile Research Journal, 91(19–20), 2357–2386. https://doi.org/10.1177/00405175211006212
[35] Tanaka, F. H. K. D. S., & Aranha, C. (2019). Data augmentation using GANs. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1904.09135
[36] Goncalves, A., Ray, P., Soper, B., Stevens, J., Coyle, L., & Sales, A. P. (2020). Generation and evaluation of synthetic patient data. BMC Medical Research Methodology, 20(1). https://doi.org/10.1186/s12874-020-00977-1
[37] Jadon, A., & Kumar, S. (2023). Leveraging Generative AI Models for Synthetic Data Generation in Healthcare: Balancing Research and Privacy. In 2023 International Conference on Smart Applications, Communications and Networking (SmartNets), Istanbul, Turkiye, pp. 1-4. https://doi.org/10.1109/smartnets58706.2023.10215825
[38] Mustak, M., Salminen, J., Mäntymäki, M., Rahman, A., & Dwivedi, Y. K. (2022). Deepfakes: Deceptions, mitigations, and opportunities. Journal of Business Research, 154, 113368. https://doi.org/10.1016/j.jbusres.2022.113368
[39] Karnouskos, S. (2020). Artificial intelligence in digital media: the era of Deepfakes. IEEE Transactions on Technology and Society, 1(3), 138–147. https://doi.org/10.1109/tts.2020.3001312
[40] Fu, R., Huang, Y., & Singh, P. V. (2020). AI and Algorithmic Bias: Source, Detection, Mitigation and Implications. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3681517
[41] Van Giffen, B., Herhausen, D., & Fahse, T. (2022). Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. Journal of Business Research, 144, 93–106. https://doi.org/10.1016/j.jbusres.2022.01.076
[42] Chin, M. H., Afsar-Manesh, N., Bierman, A. S., Chang, C., Colón-Rodríguez, C. J., Dullabh, P., Duran, D. G., Fair, M., Hernandez-Boussard, T., Hightower, M., Jain, A., Jordan, W. B., Konya, S., Moore, R. H., Moore, T. T., Rodriguez, R., Shaheen, G., Snyder, L. P., Srinivasan, M., . . . Ohno-Machado, L. (2023). Guiding principles to address the impact of algorithm bias on racial and ethnic disparities in health and health care. JAMA Network Open, 6(12), e2345050. https://doi.org/10.1001/jamanetworkopen.2023.45050
[43] Pfeiffer, J., Gutschow, J., Haas, C., Möslein, F., Maspfuhl, O., Borgers, F., & Alpsancar, S. (2023). Algorithmic fairness in AI. Business & Information Systems Engineering, 65(2), 209–222. https://doi.org/10.1007/s12599-023-00787-x

[44] Strubell, E., Ganesh, A., & McCallum, A. (2020). Energy and policy considerations for modern deep learning research. Proceedings of the AAAI Conference on Artificial Intelligence, 34(09), 13693–13696. https://doi.org/10.1609/aaai.v34i09.7123
[45] Bolón-Canedo, V., Morán-Fernández, L., Cancela, B., & Alonso-Betanzos, A. (2024). A review of green artificial intelligence: Towards a more sustainable future. Neurocomputing, 599, 128096. https://doi.org/10.1016/j.neucom.2024.128096
[46] Maleki, N., Padmanabhan, B., & Dutta, K. (2024). AI Hallucinations: A Misnomer Worth Clarifying. In 2024 IEEE Conference on Artificial Intelligence (CAI), Singapore, Singapore, pp. 133-138. https://doi.org/10.1109/cai59869.2024.00033
[47] Reddy, G. P., Pavan Kumar, Y. V. P., & Prakash, K. P. (2024). Hallucinations in Large Language Models (LLMs). In 2024 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream), Vilnius, Lithuania, pp. 1-6 (pp. 1–6). https://doi.org/10.1109/estream61684.2024.10542617
[48] Hanna, E., & Levic, A. (2023). Comparative Analysis of Language Models: hallucinations in ChatGPT: Prompt Study [BA thesis, Linnaeus University]. https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1764165&dswid=6606
[49] Amatriain, X. (2024). Measuring and mitigating hallucinations in large language models: a multifaceted approach. In amatria.in. Retrieved December 7, 2024, from https://amatria.in/blog/images/Mitigating_Hallucinations.pdf
[50] Oelschlager, R. (2024). Evaluating the Impact of Hallucinations on User Trust and Satisfaction in LLM-based Systems [BA Thesis, Linnaeus University]. https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1870904&dswid=-4068
[51] Colasacco, C. J., & Born, H. L. (2024). A case of artificial intelligence chatbot hallucination. JAMA Otolaryngology–Head & Neck Surgery, 150(6), 457. https://doi.org/10.1001/jamaoto.2024.0428
[52] Ahmadi, A. (2024). Unravelling the mysteries of hallucination in large Language models: Strategies for Precision in Artificial Intelligence Language Generation. Asian Journal of Computer Science and Technology, 13(1), 1–10. https://doi.org/10.70112/ajcst-2024.13.1.4144
[53] Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. ACM Computing Surveys, 41(3), 1–52. https://doi.org/10.1145/1541880.1541883
[54] Abedjan, Z., Chu, X., Deng, D., Fernandez, R. C., Ilyas, I. F., Ouzzani, M., Papotti, P., Stonebraker, M., & Tang, N. (2016). Detecting data errors. Proceedings of the VLDB Endowment, 9(12), 993–1004. https://doi.org/10.14778/2994509.2994518
[55] Arndt, S., & Woolson, R. W. (1993). Assessment of data quality: errors of measurement and errors of process. Biometrical Journal, 35(3), 315–324. https://doi.org/10.1002/bimj.4710350307
[56] Fisher, C. W., Chengalur-Smith, I., & Ballou, D. P. (2003). The impact of experience and time on the use of data quality information in decision making. Information Systems Research, 14(2), 170–188. https://doi.org/10.1287/isre.14.2.170.16017
[57] Redman, T. C. (1998). The impact of poor data quality on the typical enterprise. Communications of the ACM, 41(2), 79–82. https://doi.org/10.1145/269012.269025
[58] Chengalur-Smith, I., Ballou, D., & Pazer, H. (1999). The impact of data quality information on decision making: an exploratory analysis. IEEE Transactions on Knowledge and Data Engineering, 11(6), 853–864. https://doi.org/10.1109/69.824597
[59] Bruls, E. (1995). Quality and reliability impact of defect data analysis. IEEE Transactions on Semiconductor Manufacturing, 8(2), 121–129. https://doi.org/10.1109/66.382275
[60] Alruhaymi, A. Z., & Kim, C. J. (2021). Study on the missing data mechanisms and imputation methods. Open Journal of Statistics, 11(04), 477–492. https://doi.org/10.4236/ojs.2021.114030
[61] García, S., Luengo, J., & Herrera, F. (2014). Data preprocessing in data mining. In Intelligent systems reference library. Springer Cham. https://doi.org/10.1007/978-3-319-10247-4
[62] Ganti, V., & Sarma, A. D. (2013). Data cleaning: A Practical Perspective (1st ed.). Springer Cham. https://doi.org/10.1007/978-3-031-01897-8
[63] Haider, S. N., Zhao, Q., & Meran, B. K. (2020). Automated data cleaning for data centers: A case study [Paper]. IEEE. https://doi.org/10.23919/ccc50068.2020.9189357
[64] Mahdavi, M., Neutatz, F., Visengeriyeva, L., & Abedjan, Z. (2019). Towards Automated Data Cleaning Workflows [Conference Paper]. LWDA 2019, Berlin, Germany. https://www.researchgate.net/publication/335136628_Towards_Automated_Data_Cleaning_Workflows
[65] 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 & 2. https://personales.upv.es/thinkmind/dl/journals/soft/soft_v10_n12_2017/soft_v10_n12_2017_1.pdf

[66] Bogatu, A., Paton, N. W., Fernandes, A. a. A., & Koehler, M. (2018). Towards automatic data format transformations: data wrangling at scale. The Computer Journal, 62(7), 1044–1060. https://doi.org/10.1093/comjnl/bxy118

[67] Sun, Y., Li, J., Xu, Y., Zhang, T., & Wang, X. (2023). Deep learning versus conventional methods for missing data imputation: A review and comparative study. Expert Systems With Applications, 227, 120201. https://doi.org/10.1016/j.eswa.2023.120201

[68] Samad, M. D., Abrar, S., & Diawara, N. (2022). Missing value estimation using clustering and deep learning within multiple imputation framework. Knowledge-Based Systems, 249, 108968. https://doi.org/10.1016/j.knosys.2022.108968

[69] Richman, M. B., Trafalis, T. B., & Adrianto, I. (2009). Missing data imputation through machine learning algorithms. In S. E. Haupt, A. Pasini, & C. Marzban (Eds.), Artificial Intelligence Methods in the Environmental Sciences (pp. 153–169). Springer. https://doi.org/10.1007/978-1-4020-9119-3_7

[70] Liu, M., Li, S., Yuan, H., Ong, M. E. H., Ning, Y., Xie, F., Saffari, S. E., Shang, Y., Volovici, V., Chakraborty, B., & Liu, N. (2023). Handling missing values in healthcare data: A systematic review of deep learning-based imputation techniques. Artificial Intelligence in Medicine, 142, 102587. https://doi.org/10.1016/j.artmed.2023.102587

[71] Abedjan, Z. (2022). Enabling data-centric AI through data quality management and data literacy. It - Information Technology, 64(1–2), 67–70. https://doi.org/10.1515/itit-2021-0048

[72] Hassan, M. M., Knipper, A., & Santu, S. K. K. (2023). ChatGPT as your Personal Data Scientist. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2305.13657

[73] Jaimovitch-López, G., Ferri, C., Hernández-Orallo, J., Martínez-Plumed, F., & Ramírez-Quintana, M. J. (2022). Can language models automate data wrangling? Machine Learning, 112(6), 2053–2082. https://doi.org/10.1007/s10994-022-06259-9

[74] Ghazzai, S., Grigori, D., Benatallah, B., & Rebai, R. (2024). Harnessing GPT for Data Transformation Tasks. 2024 IEEE International Conference on Web Services (ICWS), Shenzhen, China. https://doi.org/10.1109/icws62655.2024.00160

[75] Zhang, Z., Groth, P., Calixto, I., & Schelter, S. (2024). Directions Towards Efficient and Automated Data Wrangling with Large Language Models (pp. 301–304). 2024 IEEE 40th International Conference on Data Engineering Workshops (ICDEW). https://doi.org/10.1109/icdew61823.2024.00044

[76] Zhang, H., Dong, Y., Xiao, C., & Oyamada, M. (2023). Large language models as data preprocessors. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2308.16361

[77] Naeem, Z. A., Ahmad, M. S., Eltabakh, M., Ouzzani, M., & Tang, N. (2023). RetClean: Retrieval-Based data cleaning using foundation models and data lakes. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2303.16909

[78] Liu, L., Hasegawa, S., Sampat, S. K., Xenochristou, M., Chen, W., Kato, T., Kakibuchi, T., & Asai, T. (2024). AutoDW: Automatic Data Wrangling Leveraging Large Language Models (pp. 2041–2052). ASE '24: Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering. https://doi.org/10.1145/3691620.3695267

[79] Lappin, S. (2023). Assessing the strengths and weaknesses of large language models. Journal of Logic Language and Information, 33(1), 9–20. https://doi.org/10.1007/s10849-023-09409-x

[80] Hadi, M. U., Tashi, Q. A., Qureshi, R., Shah, A., Muneer, A., Irfan, M., Zafar, A., Shaikh, M. B., Akhtar, N., Wu, J., & Mirjalili, S. (2023). Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects. TechRxiv. https://doi.org/10.36227/techrxiv.23589741.v4

[81] Mondal, B. (2019). Artificial intelligence: state of the art. In V. Balas, R. Kumar, & R. Srivastava (Eds.), Recent Trends and Advances in Artificial Intelligence and Internet of Things. Intelligent Systems Reference Library (Vol. 172, pp. 389–425). Springer. https://doi.org/10.1007/978-3-030-32644-9_32

[82] Ott, S., Barbosa-Silva, A., Blagec, K., Brauner, J., & Samwald, M. (2022). Mapping global dynamics of benchmark creation and saturation in artificial intelligence. Nature Communications, 13(1), 6793. https://doi.org/10.1038/s41467-022-34591-0

[83] Kiela, D., Bartolo, M., Nie, Y., Kaushik, D., Geiger, A., Wu, Z., Vidgen, B., Prasad, G., Singh, A., Ringshia, P., Ma, Z., Thrush, T., Riedel, S., Waseem, Z., Stenetorp, P., Jia, R., Bansal, M., Potts, C., & Williams, A. (2021). DynaBench: Rethinking Benchmarking in NLP. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 4110–4124. https://doi.org/10.18653/v1/2021.naacl-main.324

[84] Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1811.12808

[85] Sallam, M., Khalil, R., & Sallam, M. (2024). Benchmarking Generative AI: a call for establishing a comprehensive framework and a generative AIQ test. Mesopotamian Journal of Artificial Intelligence in Healthcare, 2024, 69–75. https://doi.org/10.58496/mjaih/2024/010

[86] Yao, R., Ye, Y., Zhang, J., Li, S., & Wu, O. (2022). Exploring developments of the AI field from the perspective of methods, datasets, and metrics. Information Processing & Management, 60(2), 103157. https://doi.org/10.1016/j.ipm.2022.103157
[87] Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. Scientific Reports, 14, 6086. https://doi.org/10.1038/s41598-024-56706-x
[88] Arshi, O., & Chaudhary, A. (2024). Overview of Artificial General Intelligence (AGI). In S. El Hajjami, K. Kaushik, & I. U. Khan (Eds.), Artificial General Intelligence (AGI) Security. Advanced Technologies and Societal Change. Springer, Singapore. https://doi.org/10.1007/978-981-97-3222-7_1
[89] Raji, I. D., Bender, E. M., Paullada, A., Denton, E., & Hanna, A. (2021). AI and the Everything in the Whole Wide World Benchmark. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2111.15366
[90] Sai, A. B., Mohankumar, A. K., & Khapra, M. M. (2022). A survey of evaluation metrics used for NLG systems. ACM Computing Surveys, 55(2), 1–39. https://doi.org/10.1145/3485766
[91] Khapra, M. M., & Sai, A. B. (2021). A Tutorial on Evaluation Metrics used in Natural Language Generation. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials, pages 15–19, Online (pp. 15–19). https://doi.org/10.18653/v1/2021.naacl-tutorials.4
[92] Reuel, A., Hardy, A., Smith, C., Lamparth, M., Hardy, M., & Kochenderfer, M. J. (2024). BetterBench: Assessing AI benchmarks, uncovering issues, and establishing best practices. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2411.12990
[93] Gao, W., Luo, C., Wang, L., Xiong, X., Chen, J., Hao, T., Jiang, Z., Fan, F., Du, M., Huang, Y., Zhang, F., Wen, X., Zheng, C., He, X., Dai, J., Ye, H., Cao, Z., Jia, Z., Zhan, K., . . . Zhan, J. (2019). AIBench: Towards scalable and comprehensive datacenter AI benchmarking. In C. Zheng & J. Zhan (Eds.), Benchmarking, Measuring, and Optimizing. Bench 2018. Lecture Notes in Computer Science (Vol. 11459, pp. 3–9). https://doi.org/10.1007/978-3-030-32813-9_1
[94] Martinez-Plumed, F., & Hernandez-Orallo, J. (2018). Dual indicators to analyze AI benchmarks: difficulty, discrimination, ability, and generality. IEEE Transactions on Games, 12(2), 121–131. https://doi.org/10.1109/tg.2018.2883773
[95] Fister, I., Brest, J., Iglesias, A., Galvez, A., Deb, S., & Fister, I. (2021). On selection of a benchmark by determining the algorithms' qualities. IEEE Access, 9, 51166–51178. https://doi.org/10.1109/access.2021.3058285
[96] Bourrasset, C., Boillod-Cerneux, F., Sauge, L., Deldossi, M., Wellenreiter, F., Bordawekar, R., Malaika, S., Broyelle, J., West, M., & Belgodere, B. (2019). Requirements for an enterprise AI benchmark. In R. Nambiar & M. Poess (Eds.), Performance Evaluation and Benchmarking for the Era of Artificial Intelligence. TPCTC 2018. Lecture Notes in Computer Science (Vol. 11135, pp. 71–81). Springer. https://doi.org/10.1007/978-3-030-11404-6_6
[97] Vogelsgesang, A., Haubenschild, M., Finis, J., Kemper, A., Leis, V., Muehlbauer, T., Neumann, T., & Then, M. (2018). Get Real: How Benchmarks Fail to Represent the Real World. In DBTest '18: Proceedings of the Workshop on Testing Database Systems. Association for Computing Machinery. https://doi.org/10.1145/3209950.3209952
[98] Bandi, A., Adapa, P. V. S. R., & Kuchi, Y. E. V. P. K. (2023). The power of Generative AI: a review of requirements, models, Input–Output formats, evaluation metrics, and challenges. Future Internet, 15(8), 260. https://doi.org/10.3390/fi15080260
[99] Andrianova, A., Simonov, M., Perets, D., Margarit, A., Serebryakova, D., Bogdanov, Y., Budennyy, S., Volkov, N., Tsanda, A., & Bukharev, A. (2018). Application of machine learning for oilfield data quality improvement. Day 2 Tue, October 04, 2022. https://doi.org/10.2118/191601-18rptc-ms
[100] Bruni, R. (2005). Error correction for massive datasets. Optimization Methods & Software, 20(2–3), 297–316. https://doi.org/10.1080/10556780512331318281
[101] Guo, X., & Chen, Y. (2024). Generative AI for synthetic data generation: methods, challenges and the future. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2403.04190
[102] Liu, S., Chen, J., Feng, Y., Xie, Z., Pan, T., & Xie, J. (2024). Generative artificial intelligence and data augmentation for prognostic and health management: Taxonomy, progress, and prospects. Expert Systems With Applications, 255, 124511. https://doi.org/10.1016/j.eswa.2024.124511
[103] Liu, L., Meng, J., & Yang, Y. (2024). LLM technologies and information search. Journal of Economy and Technology, 2, 269–277. https://doi.org/10.1016/j.ject.2024.08.007
[104] Hersh, W. R. (2023). Search Still matters: Information Retrieval in the era of Generative AI. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2311.18550
[105] Wang, P., He, Y., Shea, R., Wang, J., & Wu, E. (2018). Deeper: A Data Enrichment System Powered by Deep Web. Proceedings of the 2022 International Conference on Management of Data, 1801–1804. https://doi.org/10.1145/3183713.3193569

[106] Azad, S. A., Wasimi, S., & Ali, A. S. (2018). Business Data Enrichment: Issues and Challenges. In 2018 5th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE) (pp. 98–102). https://doi.org/10.1109/apwconcse.2018.00024
[107] Einy, Y., Milo, T., & Novgorodov, S. (2024). Cost-Effective LLM Utilization for Machine Learning Tasks over Tabular Data. In GUIDE-AI '24: Proceedings of the Conference on Governance, Understanding and Integration of Data for Effective and Responsible AI. https://doi.org/10.1145/3665601.3669848
[108] Kusano, G. (2024). GA-Tag: Data Enrichment with an Automatic Tagging System Utilizing Large Language Models. In 2024 IEEE 40th International Conference on Data Engineering (ICDE) (pp. 5397–5400). IEEE. https://doi.org/10.1109/icde60146.2024.00412
[109] Avogadro, R. (2024). Semantic Enrichment of Tabular Data with Machine Learning Techniques [PhD Dissertation, University of Milano-Bicocca]. https://hdl.handle.net/10281/465138
[110] Kasneci, G., & Kasneci, E. (2024). Enriching Tabular Data with Contextual LLM Embeddings: A Comprehensive Ablation Study for Ensemble Classifiers. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2411.01645
[111] Ispayev, S., Khamzin, N., Revkova, O., Uanassov, B., & Sison, M. (2024). Asset and Materials Data Enrichment Automation using Computer Vision and Large Language Model. In Paper presented at the SPE Caspian Technical Conference and Exhibition, Atyrau, Kazakhstan. https://doi.org/10.2118/223468-ms
[112] Cirillo, S., Desiato, D., Polese, G., & Sebillo, M. (2024). Augmenting Anonymized Data with AI: Exploring the Feasibility and Limitations of Large Language Models in Data Enrichment. In ITADATA2024: The 3rd Italian Conference on Big Data and Data Science.
[113] Gomadam, K., Yeh, P. Z., Verma, K., & Miller, J. A. (2012). Data Enrichment Using Web APIs. In 2012 IEEE First International Conference on Services Economics, Honolulu, HI, USA (pp. 46–53). IEEE. https://doi.org/10.1109/se.2012.17
[114] Weerasinghe, M. (2024). Enhancing Web Scraping with Artificial Intelligence: A Review. In 4th Research Symposium of Faculty of Computing 2024, General Sir John Kotelawala Defence University. https://www.researchgate.net/publication/379024314_Enhancing_Web_Scraping_with_Artificial_Intelligence_A_Review
[115] Gomadam, K., Yeh, P. Z., & Verma, K. (2012). Data Enrichment Using Data Sources on the Web (SS-12-04). Accenture Technology Labs. Retrieved November 1, 2024, from https://cdn.aaai.org/ocs/4336/4336-19503-1-PB.pdf
[116] Hawkins, W., & Mittelstadt, B. (2023). The ethical ambiguity of AI data enrichment: Measuring gaps in research ethics norms and practices. 2022 ACM Conference on Fairness, Accountability, and Transparency, 261–270. https://doi.org/10.1145/3593013.3593995
[117] Javaid, H. A. (2024). AI-Driven Predictive Analytics in Finance: Transforming Risk Assessment and Decision-Making. Advances in Computer Sciences, 7(1). https://academicpinnacle.com/index.php/acs/article/view/204/216
[118] Adesina, N. a. A., Iyelolu, N. T. V., & Paul, N. P. O. (2024). Leveraging predictive analytics for strategic decision-making: Enhancing business performance through data-driven insights. World Journal of Advanced Research and Reviews, 22(3), 1927–1934. https://doi.org/10.30574/wjarr.2024.22.3.1961
[119] Roselli, D., Matthews, J., & Talagala, N. (2019). Managing Bias in AI. Companion Proceedings of the 2019 World Wide Web Conference, 539–544. https://doi.org/10.1145/3308560.3317590
[120] Chen, J., Storchan, V., & Kurshan, E. (2021). Beyond Fairness Metrics: Roadblocks and challenges for ethical AI in practice. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2108.06217
[121] Martin, K. D., & Zimmermann, J. (2024). Artificial intelligence and its implications for data privacy. Current Opinion in Psychology, 58, 101829. https://doi.org/10.1016/j.copsyc.2024.101829
[122] Rai, A. (2020). Explainable AI: from black box to glass box. Journal of the Academy of Marketing Science, 48(1), 137–141. https://doi.org/10.1007/s11747-019-00710-5
[123] Naidu, G., Zuva, T., & Sibanda, E. M. (2023). A review of evaluation metrics in Machine learning Algorithms. In R. Silhavy & P. Silhavy (Eds.), Artificial Intelligence Application in Networks and Systems. CSOC 2023. Lecture Notes in Networks and Systems (Vol. 724, pp. 15–25). Springer. https://doi.org/10.1007/978-3-031-35314-7_2
[124] Rawte, V., Chakraborty, S., Pathak, A., Sarkar, A., Tonmoy, S. M. T. I., Chadha, A., Sheth, A. P., & Das, A. (2023). The troubling emergence of hallucination in large language models -- an extensive definition, quantification, and prescriptive remediations. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2310.04988
[125] Nananukul, N., & Kejriwal, M. (2024). HALO: an ontology for representing and categorizing hallucinations in large language models. In Disruptive Technologies in Information Sciences VIII (Vol. 13058). https://doi.org/10.1117/12.3014048

[126] Bahrini, A., Khamoshifar, M., Abbasimehr, H., Riggs, R. J., Esmaeili, M., Majdabadkohne, R. M., & Pasehvar, M. (2023). ChatGPT: Applications, Opportunities, and Threats. In 2023 Systems and Information Engineering Design Symposium (SIEDS), Charlottesville, VA, USA. https://doi.org/10.1109/sieds58326.2023.10137850
[127] 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
[128] Morgan, D. L. (2023). Exploring the use of artificial intelligence for qualitative data analysis: the case of ChatGPT. International Journal of Qualitative Methods, 22. https://doi.org/10.1177/16094069231211248
[129] Blagec, K., Dorffner, G., Moradi, M., & Samwald, M. (2020). A critical analysis of metrics used for measuring progress in artificial intelligence. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2008.02577
[130] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: a Multi-Task benchmark and analysis platform for natural language understanding. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1804.07461
[131] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision, 115(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y
[132] Moor, J. H. (2003). The Turing Test. In Studies in cognitive systems. https://doi.org/10.1007/978-94-010-0105-2
[133] French, R. M. (2000). The Turing Test: the first 50 years. Trends in Cognitive Sciences, 4(3), 115–122. https://doi.org/10.1016/s1364-6613(00)01453-4
[134] Saygin, A. P., Cicekli, I., & Akman, V. (2000). Turing Test: 50 Years Later. Minds and Machines, 10, 463–518. https://doi.org/10.1023/a:1011288000451
[135] Newell, A., & Simon, H. A. (1995). GPS, a program that simulates human thought. In Computation & intelligence: collected readings (pp. 415–428).
[136] Culatta, R. (2018, November 30). General Problem Solver (A. Newell & H. Simon). InstructionalDesign.org. https://www.instructionaldesign.org/theories/general-problem-solver/
[137] Vakili, M., Ghamsari, M., & Rezaei, M. (2020). Performance analysis and comparison of machine and deep learning algorithms for IoT data classification. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2001.09636
[138] Khan, M. M. R., Arif, R. B., Siddique, M. a. B., & Oishe, M. R. (2018). Study and Observation of the Variation of Accuracies of KNN, SVM, LMNN, ENN Algorithms on Eleven Different Datasets from UCI Machine Learning Repository. In 2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT), Dhaka, Bangladesh, pp. 124-129 (pp. 124–129). https://doi.org/10.1109/ceeict.2018.8628041
[139] UC Irvine Machine Learning Repository. (2023). Retrieved November 4, 2024, from https://archive.ics.uci.edu/
[140] Jones, K. S., & Galliers, J. R. (2005). Evaluating natural language processing systems: An Analysis and Review. In Lecture notes in computer science (1st ed.). Springer Berlin. https://doi.org/10.1007/BFb0027470
[141] Kasturi, R., Goldgof, D., Soundararajan, P., Manohar, V., Garofolo, J., Bowers, R., Boonstra, M., Korzhova, V., & Zhang, N. J. (2009). Framework for Performance Evaluation of face, text, and vehicle detection and tracking in video: Data, Metrics, and Protocol. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(2), 319–336. https://doi.org/10.1109/tpami.2008.57
[142] Nassif, A. B., Shahin, I., Attili, I., Azzeh, M., & Shaalan, K. (2019). Speech Recognition Using Deep Neural Networks: A Systematic review. IEEE Access, 7, 19143–19165. https://doi.org/10.1109/access.2019.2896880
[143] Chowdhury, G. (2007). TREC: Experiment and Evaluation in Information retrieval. Online Information Review, 31(5), 717–718. https://doi.org/10.1108/14684520710832478
[144] O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1511.08458
[145] Yao, G., Lei, T., & Zhong, J. (2018). A review of Convolutional-Neural-Network-based action recognition. Pattern Recognition Letters, 118, 14–22. https://doi.org/10.1016/j.patrec.2018.05.018
[146] Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., & Chen, T. (2017). Recent advances in convolutional neural networks. Pattern Recognition, 77, 354–377. https://doi.org/10.1016/j.patcog.2017.10.013
[147] Schmidt, R. M. (2019). Recurrent Neural Networks (RNNs): A gentle Introduction and Overview. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1912.05911

[148] Li, S. (2024). Development of recurrent neural networks and its applications to activity recognition [Thesis, University of Wollongong]. In figshare. https://hdl.handle.net/10779/uow.27667512.v1
[149] Brunner, F. (2019). Mastering the game of Go with deep neural networks and tree search (Silver et al., 2016) (Artificial Intelligence for Games Seminar Report). The Heidelberg Collaboratory for Image Processing (HCI). Retrieved November 13, 2024, from https://hci.iwr.uni-heidelberg.de/system/files/private/downloads/36349047/report_florian_brunner.pdf
[150] Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., Van Den Driessche, G., Graepel, T., & Hassabis, D. (2017). Mastering the game of Go without human knowledge. Nature, 550(7676), 354–359. https://doi.org/10.1038/nature24270
[151] Tian, Y., Ma, J., Gong, Q., Sengupta, S., Chen, Z., Pinkerton, J., & Zitnick, L. (2019). ELF OpenGo: an analysis and open reimplementation of AlphaZero. In Proceedings of the 36th International Conference on Machine Learning, PMLR 97:6244-6253.
[152] Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019). SuperGLUE: a stickier benchmark for General-Purpose Language Understanding Systems. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1905.00537
[153] Jiang, D., Ku, M., Li, T., Ni, Y., Sun, S., Fan, R., & Chen, W. (2024). GenAI Arena: an open evaluation platform for generative models. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2406.04485
[154] Li, B., Lin, Z., Pathak, D., Li, J., Fei, Y., Wu, K., Ling, T., Xia, X., Zhang, P., Neubig, G., & Ramanan, D. (2024). GenAI-Bench: Evaluating and Improving Compositional Text-to-Visual Generation. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2406.13743
[155] Betzalel, E., Penso, C., Navon, A., & Fetaya, E. (2022). A study on the evaluation of generative models. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2206.10935
[156] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: A method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318. https://aclanthology.org/P02-1040.pdf
[157] Lin, C. Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out, pages 74–81, Barcelona, Spain. Association for Computational Linguistics. https://aclanthology.org/W04-1013/
[158] Lin, C., & Hovy, E. (2003). Automatic evaluation of summaries using N-gram co-occurrence statistics. In NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, pp. 71 - 78 (Vol. 1, pp. 71–78). https://doi.org/10.3115/1073445.1073465
[159] Yu, Y., Zhang, W., & Deng, Y. (2021). Frechet Inception Distance (FID) for Evaluating GANs. https://www.researchgate.net/publication/354269184_Frechet_Inception_Distance_FID_for_Evaluating_GANs
[160] Obukhov, A., & Krasnyanskiy, M. (2020). Quality Assessment Method for GAN Based on Modified Metrics Inception Score and Fréchet Inception Distance. In R. Silhavy, P. Silhavy, & Z. Prokopova (Eds.), Software Engineering Perspectives in Intelligent Systems. Advances in Intelligent Systems and Computing (pp. 102–114). Springer. https://doi.org/10.1007/978-3-030-63322-6_8
[161] De Deijn, R., Batra, A., Koch, B., Mansoor, N., & Makkena, H. (2024). Reviewing Fid and Sid Metrics on Generative Adversarial Networks. arXiv (Cornell University), 111–124. https://doi.org/10.5121/csit.2024.140208
[162] Barratt, S., & Sharma, R. (2018). A note on the inception score. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1801.01973
[163] Arabboev, M., Begmatov, S., Rikhsivoev, M., Nosirov, K., & Saydiakbarov, S. (2024). comprehensive review of image super-resolution metrics: classical and AI-based approaches. ACTA IMEKO, 13(1), 1–8. https://doi.org/10.21014/actaimeko.v13i1.1679
[164] Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2024). A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology, 15(3), 39. https://doi.org/10.1145/3641289
[165] Wang, X., Jiang, L., Hernandez-Orallo, J., Sun, L., Stillwell, D., Luo, F., & Xie, X. (2023). Evaluating General-Purpose AI with Psychometrics. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2310.16379
[166] Li, L., Chen, G., Shi, H., Xiao, J., & Chen, L. (2024). A survey on Multimodal Benchmarks: In the era of large AI models. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2409.18142
[167] Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A., MD, Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., . . . Wu, Z. (2022). Beyond the Imitation Game:

Quantifying and extrapolating the capabilities of language models. Transactions on Machine Learning Research. https://doi.org/10.48550/arxiv.2206.04615
[168] McIntosh, T. R., Susnjak, T., Liu, T., Watters, P., & Halgamuge, M. N. (2024). Inadequacies of large language model benchmarks in the era of generative artificial intelligence. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2402.09880
[169] Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., Li, T., Ku, M., Wang, K., Zhuang, A., Fan, R., Yue, X., & Chen, W. (2024). MMLU-Pro: a more robust and challenging Multi-Task Language Understanding benchmark. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2406.01574
[170] Bommasani, R., Liang, P., & Lee, T. (2023). Holistic evaluation of language models. Annals of the New York Academy of Sciences, 1525(1), 140–146. https://doi.org/10.1111/nyas.15007
[171] Stanford Vision Lab. (2020). ImageNet Large Scale Visual Recognition Challenge (ILSVRC). ImageNet. Retrieved December 3, 2024, from https://www.image-net.org/challenges/LSVRC/
[172] Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., C, V. E. B., Awwal, A. a. S., & Asari, V. K. (2018). The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1803.01164
[173] Muhammad, U., Wang, W., Chattha, S. P., & Ali, S. (2018). Pre-trained VGGNet Architecture for Remote-Sensing Image Scene Classification. 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 1622–1627. https://doi.org/10.1109/ICPR.2018.8545591
[174] Khan, R. U., Zhang, X., Kumar, R., & Aboagye, E. O. (2018). Evaluating the Performance of ResNet Model Based on Image Recognition. Proceedings of the 2018 International Conference on Computing and Artificial Intelligence, 86–90. https://doi.org/10.1145/3194452.3194461
[175] Alaparthi, S., & Mishra, M. (2020). Bidirectional Encoder Representations from Transformers (BERT): A sentiment analysis odyssey. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2007.01127
[176] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). ROBERTA: A robustly optimized BERT pretraining approach. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1907.11692
[177] Yenduri, G., Ramalingam, M., Selvi, G. C., Supriya, Y., Srivastava, G., Maddikunta, P. K. R., Raj, G. D., Jhaveri, R. H., Prabadevi, B., Wang, W., Vasilakos, A. V., & Gadekallu, T. R. (2024). GPT (Generative Pre-Trained Transformer)— a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. IEEE Access, 12, 54608–54649. https://doi.org/10.1109/access.2024.3389497
[178] Chatzikoumi, E. (2019). How to evaluate machine translation: A review of automated and human metrics. Natural Language Engineering, 26(2), 137–161. https://doi.org/10.1017/s1351324919000469
[179] Jayasumana, S., Ramalingam, S., Veit, A., Glasner, D., Chakrabarti, A., & Kumar, S. (2024). Rethinking FID: Towards a Better Evaluation Metric for Image Generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9307-9315. https://doi.org/10.48550/arXiv.2401.09603
[180] Crawshaw, M. (2020). Multi-Task Learning with Deep Neural Networks: A Survey. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2009.09796
[181] Bilen, H., Rebuffi, S., & Jakab, T. (2017). Visual Domain Decathlon [Workshop]. In Part of the "PASCAL in Detail" workshop at the Conference on Computer Vision and Pattern Recognition (CVPR), 2017, Honolulu. Part of PASCAL in Detail Workshop Challenge. Conference on Computer Vision and Pattern Recognition (CVPR), 2017, Honolulu. https://www.robots.ox.ac.uk/~vgg/decathlon/#acks
[182] Rebuffi, S., Bilen, H., & Vedaldi, A. (2017). Learning multiple visual domains with residual adapters. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1705.08045
[183] Siddhant, A., Hu, J., Johnson, M., Firat, O., & Ruder, S. (2020). Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In Proceedings of the International Conference on Machine Learning (pp. 4411-4421). (Vol. 119). https://proceedings.mlr.press/v119/hu20b.html
[184] Zhang, L., & Zhang, L. (2022). Artificial Intelligence for Remote Sensing Data Analysis: A review of challenges and opportunities. IEEE Geoscience and Remote Sensing Magazine, 10(2), 270–294. https://doi.org/10.1109/mgrs.2022.3145854
[185] Pandimurugan, V., Rajaram, V., Srividhya, S., Saranya, G., & Rodrigues, P. (2024). Artificial Intelligence-based Data Wrangling Issues and Data Analytics Process for Various Domains. In Advancement of Data Processing Methods for Artificial and Computing Intelligence (pp. 151–181). River Publishers. https://doi.org/10.1201/9781032630212-9

[186] Petricek, T., Van Den Burg, G. J. J., Nazábal, A., Ceritli, T., Jiménez-Ruiz, E., & Williams, C. K. I. (2023). AI Assistants: A framework for Semi-Automated Data Wrangling. IEEE Transactions on Knowledge and Data Engineering, 35(9), 9295–9306. https://doi.org/10.1109/tkde.2022.3222538
[187] Eyuboglu, S., Karlaš, B., Ré, C., Zhang, C., & Zou, J. (2022). dcbench: a benchmark for data-centric AI systems. In Proceedings of the Sixth Workshop on Data Management for End-To-End Machine Learning (pp. 1–4). https://doi.org/10.1145/3533028.3533310
[188] Zhao, Y. (2024). Benchmarking machine learning models for quantum error correction. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2311.11167
[189] Catal, C., & Diri, B. (2009). Investigating the effect of dataset size, metrics sets, and feature selection techniques on software fault prediction problem. Information Sciences, 179(8), 1040–1058. https://doi.org/10.1016/j.ins.2008.12.001
[190] Shen, Y., Ai, X., Raj, A. G. S., John, R. J. L., & Syamkumar, M. (2024). Implications of ChatGPT for Data Science Education. In Proceedings of the 55th ACM Technical Symposium on Computer Science Education, pp. 1230 - 1236 (Vol. 1, pp. 1230–1236). https://doi.org/10.1145/3626252.3630874
[191] Ilyas, I. F., & Rekatsinas, T. (2022). Machine learning and data cleaning: Which serves the other? Journal of Data and Information Quality, 14(3), 1–11. https://doi.org/10.1145/3506712
[192] Abdelaal, M., Hammacher, C., & Schoening, H. (2023). REIN: a comprehensive benchmark framework for data cleaning methods in ML pipelines. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2302.04702
[193] Mazumder, M., Banbury, C., Yao, X., Karlaš, B., Rojas, W. G., Diamos, S., Diamos, G., He, L., Kiela, D., Jurado, D., Kanter, D., Mosquera, R., Ciro, J., Aroyo, L., Acun, B., Eyuboglu, S., Ghorbani, A., Goodman, E., Kane, T., . . . Reddi, V. J. (2022). DataPerf: Benchmarks for Data-Centric AI development. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2207.10062
[194] Ravichandran, P., Machireddy, J. R., & Rachakatla, S. K. (2022). Generative AI in Data Science: Applications in Automated Data Cleaning and Preprocessing for Machine Learning Models. Journal of Bioinformatics and Artificial Intelligence, 2(1), 129–152. https://biotechjournal.org/index.php/jbai/article/view/71
[195] Côté, P., Nikanjam, A., Ahmed, N., Humeniuk, D., & Khomh, F. (2024). Data cleaning and machine learning: a systematic literature review. Automated Software Engineering, 31(2). https://doi.org/10.1007/s10515-024-00453-w
[196] Zhang, Y., Jiang, Q., Han, X., Chen, N., Yang, Y., & Ren, K. (2024). Benchmarking data science agents. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2402.17168
[197] He, X., Zhang, Q., Jin, A., Yuan, Y., & Yiu, S. (2024). TUBench: Benchmarking Large Vision-Language Models on Trustworthiness with Unanswerable Questions. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2410.04107
[198] Handelman, G. S., Kok, H. K., Chandra, R. V., Razavi, A. H., Huang, S., Brooks, M., Lee, M. J., & Asadi, H. (2018). Peering into the black box of artificial intelligence: Evaluation Metrics of Machine Learning Methods. American Journal of Roentgenology, 212(1), 38–43. https://doi.org/10.2214/ajr.18.20224
[199] Netisopakul, P., & Taoto, U. (2023). Comparison of evaluation metrics for short story generation. IEEE Access, 11, 140253–140269. https://doi.org/10.1109/access.2023.3337095
[200] Marvin, G., Hellen, N., Jjingo, D., & Nakatumba-Nabende, J. (2024). Prompt engineering in large language models. In I. Jacob, S. Piramuthu, & P. Falkowski-Gilski (Eds.), Data Intelligence and Cognitive Informatics. ICDICI 2023. Algorithms for Intelligent Systems (pp. 387–402). Springer. https://doi.org/10.1007/978-981-99-7962-2_30
[201] Ekin, S. (2023). Prompt Engineering For ChatGPT: A Quick Guide To Techniques, Tips, And Best Practices. TechRxiv. https://doi.org/10.36227/techrxiv.22683919.v2
[202] Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2024). A Systematic survey of prompt engineering in large language Models: Techniques and applications. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2402.07927
[203] Bozkurt, A. (2024). Tell me your prompts and I will make them true: the alchemy of prompt engineering and generative AI. Open Praxis, 16(2), 111–118. https://doi.org/10.55982/openpraxis.16.2.661
[204] Moriarty, J. P. (2011). A theory of benchmarking. Benchmarking an International Journal, 18(4), 588–611. https://doi.org/10.1108/14635771111147650
[205] Thomas, R. L., & Uminsky, D. (2022). Reliance on metrics is a fundamental challenge for AI. Patterns, 3(5), 100476. https://doi.org/10.1016/j.patter.2022.100476
[206] Taleb, I., Kassabi, H. T. E., Serhani, M. A., Dssouli, R., & Bouhaddioui, C. (2016). Big Data Quality: A Quality Dimensions Evaluation. In 2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress

(UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld), pp. 759-765. https://doi.org/10.1109/uic-atc-scalcom-cbdcom-iop-smartworld.2016.0122
[207] Sidi, F., Panahy, P. H. S., Affendey, L. S., Jabar, M. A., Ibrahim, H., & Mustapha, A. (2012). Data quality: A survey of data quality dimensions. In 2012 International Conference on Information Retrieval & Knowledge Management, Kuala Lumpur, Malaysia, pp. 300-304 (pp. 300–304). https://doi.org/10.1109/infrkm.2012.6204995
[208] Fan, W. (2012). Data Quality: Theory and practice. In International Conference on Web-Age Information Management (WAIM) (pp. 1–16). https://doi.org/10.1007/978-3-642-32281-5_1
[209] Leavy, S., O'Sullivan, B., & Siapera, E. (2020). Data, power and bias in artificial intelligence. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2008.07341
[210] Bird, S., Kenthapadi, K., Kiciman, E., & Mitchell, M. (2019). Fairness-Aware Machine Learning. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, pp. 834-835 (pp. 834–835). https://doi.org/10.1145/3289600.3291383
[211] Partelow, S. (2023). What is a framework? Understanding their purpose, value, development and use. Journal of Environmental Studies and Sciences, 13(3), 510–519. https://doi.org/10.1007/s13412-023-00833-w
[212] OpenAI. (2024). Data analysis with ChatGPT: Feature and capabilities used when working with data in ChatGPT. Retrieved December 10, 2024, from https://help.openai.com/en/articles/8437071-data-analysis-with-chatgpt
[213] Natural Resources Canada. (2024). Canadian Wind Turbine Database [Dataset]. https://open.canada.ca/data/en/dataset/79fdad93-9025-49ad-ba16-c26d718cc070
[214] Beutel, A., Chen, J., Doshi, T., Qian, H., Woodruff, A., Luu, C., Kreitmann, P., Bischof, J., & Chi, E. H. (2019). Putting Fairness Principles into Practice. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. https://doi.org/10.1145/3306618.3314234
[215] A. Majeed and S. O. Hwang, "When AI Meets Information Privacy: The adversarial role of AI in data sharing scenario," *IEEE Access*, vol. 11, pp. 76177–76195, Jan. 2023, doi: 10.1109/access.2023.3297646.
[216] Benedick, P., Robert, J., & Traon, Y. L. (2021). A systematic approach for evaluating artificial intelligence models in industrial settings. Sensors, 21(18), 6195. https://doi.org/10.3390/s21186195
[217] Zhang, X., Sun, J., Cheng, Z., & Chen, H. (2022). Research on the embedded Mathematical model of artificial intelligence measurement. 2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), 362–366. https://doi.org/10.1109/aemcse55572.2022.00079
[218] Gao, J., Tao, C., Jie, D., & Lu, S. (2019). Invited Paper: What is AI Software Testing? and Why. In 2019 IEEE International Conference on Service-Oriented System Engineering (SOSE), San Francisco, CA, USA. https://doi.org/10.1109/sose.2019.00015
[219] Tao, C., Gao, J., & Wang, T. (2019). Testing and quality Validation for AI Software–Perspectives, Issues, and Practices. IEEE Access, 7, 120164–120175. https://doi.org/10.1109/access.2019.2937107
[220] Bandi, A., Adapa, P. V. S. R., & Kuchi, Y. E. V. P. K. (2023b). The power of Generative AI: a review of requirements, models, Input–Output formats, evaluation metrics, and challenges. Future Internet, 15(8), 260. https://doi.org/10.3390/fi15080260
[221] Fischer, M., & Lanquillon, C. (2024). Evaluation of Generative AI-Assisted Software Design and Engineering: A User-Centered Approach. Artificial Intelligence in HCI. HCII 2024. Lecture Notes in Computer Science, 14734, 31–47. https://doi.org/10.1007/978-3-031-60606-9_3
[222] McCaffrey, P., Jackups, R., Seheult, J., Zaydman, M. A., Balis, U., Thaker, H. M., Rashidi, H., & Gullapalli, R. R. (2024). Evaluating Use of Generative Artificial intelligence in clinical pathology practice: opportunities and the way forward. Archives of Pathology & Laboratory Medicine, 149(2), 130–141. https://doi.org/10.5858/arpa.2024-0208-ra
[223] Ahuja, K., Hada, R., Ochieng, M., Jain, P., Diddee, H., Maina, S., Ganu, T., Segal, S., Axmed, M., Bali, K., & Sitaram, S. (2023). MEGA: Multilingual Evaluation of Generative AI. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2303.12528
[224] Lin, A., Zhu, L., Mou, W., Yuan, Z., Cheng, Q., Jiang, A., & Luo, P. (2024). Advancing generative AI in medicine: recommendations for standardized evaluation. International Journal of Surgery, 110(8), 4547–4551. https://doi.org/10.1097/js9.0000000000001583
[225] Chen, J., Zhu, L., Mou, W., Liu, Z., Cheng, Q., Lin, A., Zhang, J., & Luo, P. (2023). STAGER checklist: Standardized Testing and Assessment Guidelines for Evaluating Generative AI Reliability. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2312.10074
[226] McIntosh, T. R., Susnjak, T., Arachchilage, N., Liu, T., Watters, P., & Halgamuge, M. N. (2024). Inadequacies of large language model benchmarks in the era of generative artificial intelligence. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2402.09880

[227] Dholakia, A., Ellison, D., Hodak, M., Dutta, D., & Binnig, C. (2024). Benchmarking Generative AI Performance Requires a Holistic Approach. In Performance Evaluation and Benchmarking: 15th TPC Technology Conference, TPCTC 2023, Vancouver, BC, Canada, August 28 – September 1, 2023, Revised Selected Papers (pp. 34–43). https://doi.org/10.1007/978-3-031-68031-1_3
[228] Orzechowski, P., & Moore, J. H. (2021). Generative and reproducible benchmarks for comprehensive evaluation of machine learning classifiers. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2107.06475
[229] Hassani, H., & Silva, E. S. (2024). Predictions from Generative Artificial Intelligence Models: Towards a New Benchmark in Forecasting Practice. Information, 15(6), 291. https://doi.org/10.3390/info15060291
[230] Mousavi, R., Kitchens, B., Oliver, A., & Abbasi, A. (2024). From lexicons to Generative AI: Benchmarking data annotation in business research. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4776480
[231] Inala, J. P., Wang, C., Drucker, S., Ramos, G., Dibia, V., Riche, N., Brown, D., Marshall, D., & Gao, J. (2024). Data analysis in the era of generative AI. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2409.18475
[232] Bhatia, S., Gandhi, T., Kumar, D., & Jalote, P. (2023). Unit Test Generation using Generative AI : A Comparative Performance Analysis of Autogeneration Tools. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2312.10622
[233] Goyal, M., & Mahmoud, Q. H. (2024). A Systematic review of Synthetic data generation techniques using Generative AI. Electronics, 13(17), 3509. https://doi.org/10.3390/electronics13173509
[234] Ramzan, F., Sartori, C., Consoli, S., & Recupero, D. R. (2024). Generative Adversarial Networks for Synthetic data Generation in Finance: Evaluating statistical similarities and quality assessment. AI, 5(2), 667–685. https://doi.org/10.3390/ai5020035
[235] Shahbazian, R., & Greco, S. (2023). Generative adversarial networks assist missing data imputation: a comprehensive survey and evaluation. IEEE Access, 11, 88908–88928. https://doi.org/10.1109/access.2023.3306721
[236] Xia, G., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., Zhang, L., & Lu, X. (2017). AID: a benchmark data set for performance evaluation of aerial scene classification. IEEE Transactions on Geoscience and Remote Sensing, 55(7), 3965–3981. https://doi.org/10.1109/tgrs.2017.2685945
[237] Yang, Y., Malaviya, C., Fernandez, J., Swayamdipta, S., Bras, R. L., Wang, J., Bhagavatula, C., Choi, Y., & Downey, D. (2020). Generative Data Augmentation for Commonsense Reasoning. Findings of the Association for Computational Linguistics: EMNLP 2020, 1008–1025. https://doi.org/10.18653/v1/2020.findings-emnlp.90
[238] Gwon, Y. N., Kim, J. H., Chung, H. S., Jung, E. J., Chun, J., Lee, S., & Shim, S. R. (2024). The use of generative AI for scientific literature searches for systematic reviews: ChatGPT and Microsoft Bing AI Performance Evaluation. JMIR Medical Informatics, 12, e51187. https://doi.org/10.2196/51187
[239] Sai, A. B., Mohankumar, A. K., & Khapra, M. M. (2020). A survey of evaluation metrics used for NLG systems. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2008.12009

# Appendix A – Less Commonly Mentioned Metrics

- BLEU, ROUGE, and METEOR: Commonly used in natural language processing for evaluating the quality of text generation by comparing machine-generated text to reference human-written text.
- Mean Absolute Percentage Error (MAPE): Useful in forecasting tasks to measure the accuracy of predictions relative to the actual values.
- Wasserstein Distance and Energy Distance: Both used to compare distributions, relevant for tasks involving probability distributions or generative modeling.
- Kullback-Leibler (KL) Divergence: Measures the difference between two probability distributions, often used in tasks involving probabilistic modeling.
- Hallucination Rate: Specific to language models, this metric evaluates the rate at which a model generates incorrect or nonsensical content.
- Utility and Fidelity: Assess how useful and faithful AI outputs are to the intended purpose or original data.
- Latent Space Similarity (Ruzicka): Evaluates the similarity in the latent space representation of data, useful in generative modeling.
- Human Evaluation: Involves subjective assessments of AI output quality, commonly used in tasks where automated metrics fall short.
- Structural Similarity Index (SSIM): Often used in image processing to evaluate the perceived quality of images by measuring structural similarity.
- Matthews Correlation Coefficient (MCC): Provides a balanced measure for binary classification, even with imbalanced classes.
- Factuality: Evaluates the correctness of factual information generated by AI models, especially important in text generation.
- Explainability and System Stability: Address the transparency and reliability of AI systems in providing consistent outputs.
- Branch and Statement Coverage: Metrics used in AI systems involving code generation or program synthesis to evaluate how much of the code is covered.
- Timeliness: Measures how quickly an AI system can deliver results, significant for real-time applications.
- Grade-Based Evaluation (A, B, C, F): A more subjective metric that categorizes AI performance into different quality grades.

# Appendix B – Prompting scripts

### Script experiment 1, difficulty level 1:

You are provided with a mixed-type dataset in a .CSV file, containing multiple rows and columns. The data types include strings, dates, booleans, and arrays, among others. Your task is to thoroughly identify and correct all errors present in the dataset while adhering to the following requirements and guidelines:

Understanding the Dataset

- The dataset contains various types of errors spread across different columns. Errors may include (but are not limited to):
  Data Type Errors: Incorrect data types that do not match the expected data type for a column. Examples:
  A string instead of a date: "Text" appears in a column expected to store dates (e.g. "2023-11-01") or numeric values.
  A numeric value in a categorical column: 1234 appears in a column for customer names.
  A floating-point number where an integer is expected: 3.14 in a column expected to store whole numbers like 3.
  Data Entry Errors: Typographical mistakes or manual data input issues. Examples:
  Typographical error: "Appl", "Aplle", or "Applle" instead of "Apple".
  Transposition error: 54321 instead of the correct value 54312.
  Partial data: "New Yo" instead of "New York" in a city column.
  Duplicate Data: Identical records or partially overlapping data entries. Examples:
  Complete duplication: Two identical rows or columns.
  Partial duplication: Two records for the same person but with slight differences (e.g. "John Doe, New York, 30" and "John Doe, NY, 30").
  Duplicate primary key: A unique identifier (e.g., 12345) appearing in multiple rows.
  Structural Errors: Issues such as mismatched column counts, incorrect formatting, or inconsistent structure within rows. Examples:
  Mismatched column counts: One row has more or fewer fields than the others.
  Incorrect formatting: A header row is included as part of the data (e.g. "Name, Age, Location" appearing mid-dataset).
  Misaligned data: Entries in the wrong columns, such as a phone number in an email field.
  Inconsistent Data: Conflicting or differently formatted data representing the same information (e.g. inconsistent date formats). Examples:
  Mixed date formats: "2023-11-25", "11/25/2023", and "25-Nov-2023" in the same column.
  Inconsistent capitalization: "Apple", "APPLE", and "apple" in a product name column.
  Conflicting labels: "NY", "New York", and "N.Y." representing the same location.
  Incorrect Data Values: Values that fall outside of expected ranges or logical constraints. Examples:
  Out-of-range value: 200 for an age field where a maximum of 120 is expected.
  Logical inconsistency: A birth date of "2025-01-01" for a record created in 2024.
  Invalid identifier: A product ID 9999 out of 800 possible.
  Missing Data: Null or empty values. Examples:
  Completely empty fields: A column with numerical data with blank entries.
  Null values: purposely empty values must be left blank
  Skipped entries: Partial records with missing essential fields like names or IDs.
  Outliers and Anomalies: Data points significantly different from the rest, potentially indicating errors. Examples:
  Extreme numeric value: A salary value of 1,000,000 in a dataset where most salaries range between 30,000 and 80,000.
  Geographic outlier: A latitude value of 0,0 (Null Island) for a customer address.
  Anomalous category: "Purple Elephant" in a column for vehicle types.
  Data Integrity Violations: Issues related to relationships and referential integrity between data elements. Examples:
  Missing or Inconsistent Primary Keys: Two lines share the same line number.
  Poorly Executed Rules or Functions: for example - data column named "VAT 15%, but function contains 20%.
  References To Other Tables Are Corupted.
  Other errors

Detection and Correction

- Identification and Validation in steps:
  Step 1: Identify and fill missing values in numeric and categorical fields. Use reasonable defaults, such as the mean for numeric fields or leave datapoints empty for categorical fields if information cannot be extracted from the dataset.
  Step 2: Correct formatting inconsistencies across all fields. For example: Ensure dates follow a consistent format.
  Step 3: Remove duplicate rows or columns based on unique identifiers and summarize how many duplicates were removed.
  Step 4: Verify logical consistency across related fields.
  Step 4: Check for other errors that were not defined in the task but are known to you.
  Step 5: Check if not mistakes or hallucinations were made.
  Step 6: Export your output in requested format.

Correction Strategy:
  Correct data type mismatches through conversion or replacement (e.g. converting a string to a date format).
  Rectify data entry errors based on context or likely intended values (e.g. typos).
  Deduplicate records as appropriate, ensuring relevant information is retained.
  Address structural inconsistencies through schema validation and adjustments.
  Standardize inconsistent data formats and resolve conflicting information.
  Handle incorrect data values based on logical rules or domain-specific knowledge.
  Impute missing data where possible, using suitable strategies such as mean, median, mode, or domain-specific estimations.
  Identify and resolve outliers using statistical techniques, and justify any removal or changes made.
  Ensure data integrity and compliance with any implied relationships or constraints.
  Make a check if no interdependencies exist between datapoints and make corrections accordingly
  Ensure that numeric fields match logical constraints

Expected Output
- Output the corrected data as a .CSV file.

Tools and Methodologies
- Use any tools or methodologies that you find appropriate, such as scripting languages (Python, R, etc.) or data cleaning libraries.
- There are no restrictions on tools, but solutions must be justifiable and reasonable.

Documentation
- Provide a summary of the steps taken to identify and correct errors.
- If assumptions were made during corrections, clearly state them.
- Where possible, include before-and-after examples of major corrections.

## Script experiment 1, difficulty level 2:

You are provided with a mixed-type dataset in a .CSV file, containing multiple rows and columns. The dataset may contain various data errors that are distributed across different columns. Your task is to identify and correct all errors while adhering to the following general guidelines:

Understanding the Dataset and Errors
- The dataset contains potential issues, such as:

  Data Type Errors: Mismatches between the expected data type and the actual values (e.g. strings where dates are expected).

  Data Entry Errors: Typographical mistakes or inconsistent formatting (e.g. Typographical error: "Appl", "Aplle", or "Applle" instead of "Apple").

  Duplicate Data: Multiple entries that may require deduplication (e.g. strings where dates are expected).

  Structural and Format Errors: Issues such as inconsistent data structure, varying lengths of arrays, etc (e.g. Misaligned data - Entries in the wrong columns, such as a phone number in an email field).

  Missing Data and Null Values: Gaps in data that should be filled or handled appropriately (e.g. Completely empty fields: A column with numerical data with blank entries).

  Outliers and Inconsistent Data: Values that are statistically unusual or formatted inconsistently (e.g. Extreme numeric value: A salary value of 1,000,000 in a dataset where most salaries range between 30,000 and 80,000).

  Other errors

Detection and Correction Requirements

- Detect all errors using appropriate checks (e.g. type validation, deduplication techniques, statistical methods).
- Correct each identified error using reasonable and contextually appropriate methods (e.g. data type conversion, fixing typos, deduplication).
- Retain the original data structure: Do not introduce new columns or modify the overall shape of the data.
- Missing data should be addressed using a suitable approach, such as imputation or contextual adjustments.

Procedure in steps
- Step 1: Identify missing values and fill them appropriately.
- Step 2: Correct formatting inconsistencies, such as ensuring numeric values are rounded to two decimal places and text is consistently capitalized. Provide a summary of your corrections.
- Step 3: Remove duplicates and verify that all fields meet logical constraints
- Step 4: Check for other errors.
- Export the results.

Expected Output
- Output the corrected dataset as a .CSV file.

Tools and Methods
- There are no restrictions on the tools you may use (e.g. scripting languages, data cleaning tools). Use any approach that suits the data correction task.

Documentation
- Provide a brief explanation of key errors found and the general approach used for correction.
- Any assumptions made during corrections should be stated.

## Script experiment 1, difficulty level 3:

You are provided with a mixed-type dataset in a .CSV file, containing 50 rows and 16 columns. The dataset may contain a range of errors distributed across different columns, such as data type mismatches, data entry errors, duplicates, missing values, and outliers. Your task is to identify and correct all errors using any approach you deem appropriate. The original data structure must be preserved: no new columns should be added, and the data format must remain consistent.

Requirements:
- Detect and correct all errors in the dataset.
- Use any methods or tools you prefer to identify and resolve issues.

Output the corrected data in a .CSV file.

## Script experiment 2 task description 2.1:

Extract missing data for the columns [Province_Territory], [Commissioning], and [Total Project Capacity (MW)] from the original dataset website: https://open.canada.ca/data/en/dataset/79fdad93-9025-49ad-ba16-c26d718cc070 and the dataset that it contains. Ensure the retrieved data matches the format and structure of the provided dataset. Provide only verified entries from the source.
Return me the enhanced dataset in .CSV file.

## Script experiment 2 task description 2.2:

Crawl web-based sources to find missing data for the columns [Province_Territory], [Commissioning], and [Total Project Capacity (MW)]. Focus on reliable sources such as project-specific or government websites. Extract relevant values and ensure they are formatted consistently with the provided dataset.
Return me the enhanced dataset in .CSV file.

## Script experiment 2 task description 2.3:

Combine Dataset B with Dataset A to fill in missing values in Dataset A. Ensure that the data sequence and formatting of Dataset A are maintained. Do not alter the original structure of Dataset A during the integration process.
Return me the enhanced dataset in .CSV file.

# Appendix C – Data Difficulty Evaluation Criteria

| Rank | Data Error Type | Detection Difficulty | Resolution Difficulty | Overall Difficulty Level | Example |
|---|---|---|---|---|---|
| 1 | Data Type Errors | Easy | Moderate | Easy to Moderate | Text Instead of Numeric ["12", "45", "Thirty", "60"], Different itme formating ["2024-11-25", "11/25/2024", "25-11-2024", "InvalidDate"], Boolean Misrepresentation [True, False, "Yes", "No", 1, 0], Mixed data types [23, "NaN", "Unknown", 45], Integer Instead of Floating-Point [1, 2, 3] instead of [1.00, 2.35, 3.67] |
| 2 | Data Entry Errors | Easy to Moderate | Moderate | Moderate | Typographical Errors ["Appl", "Aplpe", or "Applle" instead of "Apple"], Transposition Errors ["Jonh" instead of "John"], Incomplete Data ["New Y" instead of "New York"], Invalid Data Format ["25-11-2024" entered as "25112024"], Punctuation and Spacing Errors |
| 3 | Duplicate Data | Moderate | Moderate | Moderate | Duplicate datapoints, Duplicate lines, Duplicate Columns |
| 4 | Structural Errors | Moderate | Moderate to Difficult | Moderate to Difficult | Misaligned Columns ["John, 1990-01-01, New York" instead of "John, New York, 1990-01-01"], Inconsistent Column Naming ["Customer Name", "Name", and "Full Name"] |
| 5 | Inconsistent Data | Moderate | Difficult | Moderate to Difficult | Inconsistent Naming Conventions ["NY", "New York", "N.Y."], Different Units of Measurement ["5 miles" and "8 km"], Mixed Capitalization ["Apple", "apple", "APPLE"], Inconsistent Categorical Labels ["Married", "M", "Single", "S"], |
| 6 | Incorrect Data Values | Moderate | Difficult | Moderate to Difficult | Out-of-Range Values [Age: 200 instead of a realistic value (e.g., 25)], Impossible Dates ["2023-13-01"], Logical Inconsistencies [A birthdate of "2023-01-01" and a recorded age of 30 in 2024], Negative Values Where They Don't Apply [Salary: "-5000"] |
| 7 | Missing Data | Moderate | Difficult | Moderate to Difficult | Completely Missing Fields or Datapoints |
| 8 | Outliers and Anomalies | Difficult | Difficult | Difficult | Extreme Values [Temperature: -100°C where others are in range -10°C and 40°C], Punctuation Outliers [100,00 instead of 100.00], Spatial Anomalies [GPS coordinate showing "0, 0"] |
| 9 | Data Integrity Violations | Difficult | Difficult | Difficult | Missing or Inconsistent Primary Keys [Two lines share the same line number], Poorly Executed Rules or Functions [for example: data column named "VAT 15%", but function contains 20%], References To Other Tables Are Corupted |