ACADEMIA ENGINEERING

Research Article
Published: 2024-11-29
https://doi.org/10.20935/AcadEng7407

Can generative AI transform data quality? a critical discussion of ChatGPT's capabilities

Otmane Azeroual[1,*]

Academic Editors: Dimitrios A. Karras and Angeles Blanco

Abstract

Data quality (DQ) is a fundamental element for the reliability and utility of data across various domains. The emergence of generative AI technologies, such as GPT-4, has introduced innovative methods for automating data cleaning, validation, and enhancement processes. This paper investigates the role of generative AI, particularly ChatGPT, in transforming data quality. We assess the effectiveness of these technologies in error identification and correction, data consistency validation, and metadata enhancement. Our study includes empirical results demonstrating how generative AI can significantly improve DQ. The findings suggest that generative AI and ChatGPT have a transformative impact on data management practices, offering new opportunities for enhancing data quality across various applications.

Keywords: *data quality (DQ); generative AI; GPT-4; ChatGPT; data cleaning; metadata enhancement*

Citation: Azeroual O. Can generative AI transform data quality? a critical discussion of ChatGPT's capabilities. *Academia Engineering* 2024;1. https://doi.org/10.20935/AcadEng7407

1. Introduction

In the contemporary data-driven landscape, the quality of data is critical for accurate decision-making, operational efficiency, and the dependability of data-dependent systems [1]. Low data quality can lead to incorrect conclusions, operational inefficiencies, and substantial risks [2]. As organizations increasingly handle vast amounts of data, ensuring their quality has become essential.

Traditional data cleaning and validation methods, though effective, are often labor-intensive and susceptible to human error [3]. These methods generally involve manual processes such as identifying and correcting inconsistencies, validating data against predefined standards, and enriching metadata. Despite diligent efforts, human involvement introduces variability and potential inaccuracies, particularly as data volume and complexity continue to grow [4].

The advent of generative AI technologies offers promising solutions to these challenges. Generative AI, exemplified by advanced interfaces like GPT-4, provides novel approaches for automating data cleaning, validation, and enhancement processes [5]. These interfaces excel in natural language processing (NLP) tasks due to their ability to understand and generate human-like text, making them particularly adept at tasks requiring contextual understanding and linguistic capabilities [6].

GPT-4, the fourth generation of the Generative Pre-trained Transformer, has shown remarkable proficiency in various NLP tasks [7]. Its capability to generate coherent and contextually relevant text enables automation in error detection, data consistency validation, and metadata enhancement [8]. Empirical studies reveal that GPT-4's application in data quality management can lead to substantial improvements.

ChatGPT, a variant of GPT-4, is optimized for conversational tasks and can interact with data dynamically and intuitively [9]. It can automatically correct metadata errors, infer missing information, and enrich data by adding relevant details [10]. Its conversational interface facilitates a more interactive and user-friendly approach to data management, making it accessible to users with varying levels of technical expertise [11].

This paper explores the potential of generative AI, with a focus on ChatGPT, in transforming data quality. We critically evaluate whether these interfaces can be relied upon to enhance data quality. This paper includes an analysis of GPT-4 and ChatGPT's effectiveness in error correction, data consistency validation, and metadata enhancement, supported by quantitative results and case studies.

The implications of this research are profound. Demonstrating that generative AI can reliably improve data quality could revolutionize data management practices, leading to higher accuracy and efficiency while reducing reliance on manual processes. Furthermore, the scalability of AI-driven solutions could enable more effective management of larger datasets, addressing the increasing demand for high-quality data.

[1]German Centre for Higher Education Research and Science Studies (DZHW), 10117 Berlin, Germany.
[*]email: azeroual@dzhw.eu

In conclusion, this paper provides a thorough evaluation of generative AI and ChatGPT's capabilities in enhancing data quality. By establishing their reliability, we aim to support the broader adoption of these technologies in data management, contributing to more accurate, efficient, and reliable data systems.

2. Background and literature review

2.1. Data quality (DQ): definition, importance, and challenges

Data quality (DQ) refers to the condition of data based on factors such as accuracy, completeness, reliability, relevance, and timeliness [12]. High-quality data are essential for various organizational activities, including decision-making, operational processes, and strategic planning [13]. Accurate and reliable data ensure that decisions are based on factual information, leading to better outcomes. They support operational efficiency by reducing the likelihood of errors and the need for rework. Additionally, many industries are subject to regulatory requirements that necessitate precise and comprehensive data reporting. High-quality data also enhance customer satisfaction by providing accurate information and timely responses to inquiries [14].

Despite its critical importance, maintaining high-quality data presents several challenges. Manual data entry can introduce errors such as typographical mistakes, misclassifications, and omissions [15]. Integrating data from multiple sources often leads to inconsistencies and duplicate records. Data can quickly become outdated, necessitating regular updates to maintain their relevance and accuracy [16]. Complex data structures, especially unstructured data like text and images, pose difficulties in standardization and validation. As data volumes grow, the task of maintaining quality becomes increasingly challenging due to the sheer amount of data that need to be processed.

2.2. Traditional data cleaning and validation methods: overview, limitations, and need for improvement

Traditional methods for data cleaning and validation typically involve a combination of manual processes and rule-based automated systems [17]. Manual review requires data stewards to inspect datasets and identify and correct errors. This process relies heavily on human expertise and is labor-intensive. Rule-based systems use predefined rules to identify anomalies and validate data. Common rules include format checks, range checks, and consistency checks. Deduplication processes identify and merge duplicate records to ensure a single version of the truth. Standardization converts data into a common format or structure to facilitate consistency and comparison.

While these traditional methods have been foundational in data quality management, they have several limitations. Manual processes are not scalable and become impractical as data volumes increase [18]. Manual review and correction are time-consuming, leading to delays in data availability. Human intervention introduces variability, with different data stewards potentially applying different standards and practices. Rule-based systems can be rigid and may not adapt well to changing data patterns or new types of errors. Traditional methods often struggle with unstructured data such as text, images, and videos, which are increasingly prevalent in modern datasets [19].

The increasing volume, variety, and velocity of data highlight the need for more advanced and scalable solutions for data quality management [20]. The limitations of traditional methods underscore the necessity for innovative approaches that can automate and enhance data cleaning and validation processes while maintaining high levels of accuracy and consistency [21].

2.3. Generative AI and GPT-4: introduction and applications in data processing

Generative AI refers to a class of artificial intelligence models designed to generate new data instances that resemble a given training dataset [22]. These models can create text, images, audio, and other types of data, making them highly versatile tools for various applications. Generative AI encompasses a broad range of models and techniques, each with its specific capabilities and use cases.

2.3.1. GPT-4 overview

One of the most prominent generative AI models is the Generative Pre-trained Transformer (GPT) series developed by OpenAI [7]. GPT-4, the fourth iteration of the GPT model, represents a significant advancement in this series. Unlike general generative AI models that might focus on different types of data or tasks, GPT-4 is specifically designed for advanced natural language understanding and generation [23]. Its architecture enables it to process and generate coherent, contextually relevant text based on large datasets. This specialization makes GPT-4 particularly effective for applications involving complex language tasks.

2.3.2. Clarification of GPT-4 and ChatGPT

It is important to differentiate between GPT-4 and its variant ChatGPT to fully understand their applications and capabilities. GPT-4 encompasses a broad range of models within the GPT framework, each optimized for different types of tasks and data processing needs. ChatGPT, a conversational variant of GPT-4, is specifically tailored for interactive dialogue and context-aware communication [7]. While GPT-4 as a general model can perform a variety of language-related tasks, ChatGPT is designed to excel in generating human-like conversational responses.

2.3.3. Applications in data processing

Generative AI, particularly GPT-4, has found numerous applications in data processing, including error correction, data validation, metadata generation, and more (see **Table 1**). The capabilities of GPT-4 are harnessed to perform several key functions:

1. **Error Identification and Correction**: GPT-4 can analyze datasets to identify and correct errors by understanding the context and providing accurate suggestions. This capability is enhanced by its ability to generate human-like text that aligns with the intended data structure.
2. **Data Validation**: GPT-4 can validate data against predefined standards, flagging inconsistencies or anomalies with a high degree of accuracy. This function is crucial for ensuring data integrity and compliance with quality standards.
3. **Metadata Generation and Enrichment**: GPT-4 can automatically generate and enrich metadata, improving data

organization and retrieval. By understanding the content, GPT-4 can create meaningful metadata that enhance data accessibility.

4. **Text Summarization and Insight Extraction**: GPT-4 excels in summarizing large volumes of text data, extracting key information, and providing insights. This ability is useful for managing and analyzing extensive datasets efficiently.
5. **Natural Language Processing (NLP) Applications**: GPT-4's advanced NLP capabilities enable applications such as sentiment analysis, language translation, and content generation [24]. These applications leverage GPT-4's proficiency in handling nuanced language tasks.

2. 3. 4. Comparison and clarification

While GPT-4 provides a broad range of capabilities in natural language understanding and generation, ChatGPT, a variant of GPT-4, is optimized for conversational interactions. This distinction highlights that GPT-4 encompasses a diverse group of language models, each suited to different tasks. ChatGPT's design focuses specifically on engaging in dialogues and providing contextually appropriate responses, differentiating it from the more generalized GPT-4 model [7].

2. 3. 5. Transformative Potential in data quality management

Both GPT-4 and ChatGPT offer transformative potential for data quality management. By automating and enhancing traditional data cleaning, validation, and metadata management processes, these AI models address the limitations of conventional methods. Organizations can achieve higher levels of data accuracy, consistency, and completeness by leveraging GPT-4's advanced text generation capabilities and ChatGPT's conversational strengths [7].

3. Generative AI for data quality enhancement

Figure 1 illustrates the comprehensive process flow of how generative AI, particularly models like GPT-4, can be utilized to enhance data quality. This process not only improves data integrity but also enhances the overall reliability and usability of the data. The process is divided into three main components, error detection and correction, data validation, and metadata enhancement, which are described below.

3. 1. Error detection and correction

Generative AI models, particularly GPT-4, have shown substantial promise in revolutionizing error detection and correction in data management. These models leverage advanced natural language processing (NLP) capabilities to identify and rectify errors by understanding the context and semantics of data. This process is crucial for maintaining high data quality, which is essential for accurate analysis and decision-making.

- **Identification of Data Errors**

GPT-4 identifies data errors by comparing each entry against an extensive training dataset and recognizing patterns that deviate from established norms. The model's ability to understand the context and structure of the data allows it to detect typographical mistakes, inconsistent formatting, and logical inconsistencies. For example, in a dataset containing date entries in multiple formats, GPT-4 can identify these inconsistencies by comparing each entry to standard date formats and analyzing the surrounding data context. Additionally, GPT-4 can detect logical errors such as mismatched or contradictory information by cross-referencing different parts of the dataset. Empirical evidence shows that GPT-4's contextual analysis can reduce error rates significantly. For instance, in a case study [25, 26], GPT-4 identified discrepancies in patient records, such as mismatched age and birthdate fields, by cross-referencing the data with other entries and external databases, leading to a 30% reduction in data entry errors [27].

- **Correction of Data Errors**

Once errors are identified, GPT-4 generates correction suggestions based on contextual understanding. For instance, if a numerical entry falls outside the expected range, the model can infer the correct value by analyzing similar entries or applying predefined business rules. The correction process involves generating potential solutions and evaluating their fit within the dataset's context. In a practical example, GPT-4 was used to enhance data quality in a healthcare dataset [25, 26], where the model identified and corrected discrepancies, resulting in improved accuracy of patient records. Quantitative results indicate a significant reduction in data entry errors and an enhancement in data accuracy [10].

3. 2. Data validation

Generative AI ensures data consistency and accuracy through advanced validation techniques. GPT-4 performs various checks to confirm that data adhere to predefined standards and formats, which is crucial for maintaining data integrity.

- **Ensuring Data Consistency**

GPT-4 validates data consistency by ensuring that all entries conform to uniform standards. This involves verifying that data entries, such as dates, follow a consistent format and that numerical values fall within acceptable ranges. In financial datasets, for example, GPT-4 ensures that all monetary values are correctly formatted and in the appropriate currency [28, 29]. By maintaining uniform data formats and logical coherence, the model enhances overall data integrity. Consistency checks can also be applied to categorical data, ensuring that all entries fall within predefined categories.

- **Ensuring Data Accuracy**

To ensure data accuracy, GPT-4 uses cross-referencing techniques to compare data entries with external sources. For example, in research databases, GPT-4 can verify author names, publication dates, and journal titles against trusted external databases such as PubMed or CrossRef [30]. This cross-referencing process helps identify and correct discrepancies, thereby enhancing the reliability of the data. A notable case study in the financial sector demonstrated GPT-4's validation capabilities [31, 32]. The model cross-referenced transaction records with external banking databases to verify accuracy, uncovering and correcting duplicated transactions

Table 1 • Summary of generative AI applications and performance metrics.

Application problem	LLM models used	Comparison between GPT versions	Use cases	Training and test data	Quantitative results	Remarks, discussions, and open problems
Error Identification and Correction	GPT-3, GPT-4	GPT-3 vs. GPT-4	Brown, T. B. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.	Diverse Text Corpora	GPT-4 outperforms GPT-3 in certain tasks	Discussion on model scaling
Data Validation	GPT-4	-	Nori, H., King, N., McKinney, S. M., Carignan, D., & Horvitz, E. (2023). Capabilities of gpt-4 on medical challenge problems. arXiv preprint arXiv:2303.13375.	Benchmark Data	High accuracy in validation	Challenges with large datasets
Metadata Generation and Enrichment	GPT-4	-	Angelici, P. Enhancing documents review through Knowledge Graphs and Large Language Models (Doctoral dissertation).	Text Corpora	Improved metadata quality	Applications in document management
Text Summarization	GPT-4	-	Kumar, J. (2023). Large language models for text summarization: A comprehensive study. Pranjana: The Journal of Management Awareness, 26(1and2), 113-124.	Summarization Datasets	Effective summarization of large texts	Comparison of summarization capabilities
NLP Applications	GPT-4, ChatGPT	GPT-4 vs. ChatGPT	Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., ... & Ge, B. (2023). Summary of chatgpt-related research and perspective towards the future of large language models. Meta-Radiology, 100017.	Diverse NLP Datasets	Advances in sentiment analysis and translation	Conversational strengths of ChatGPT

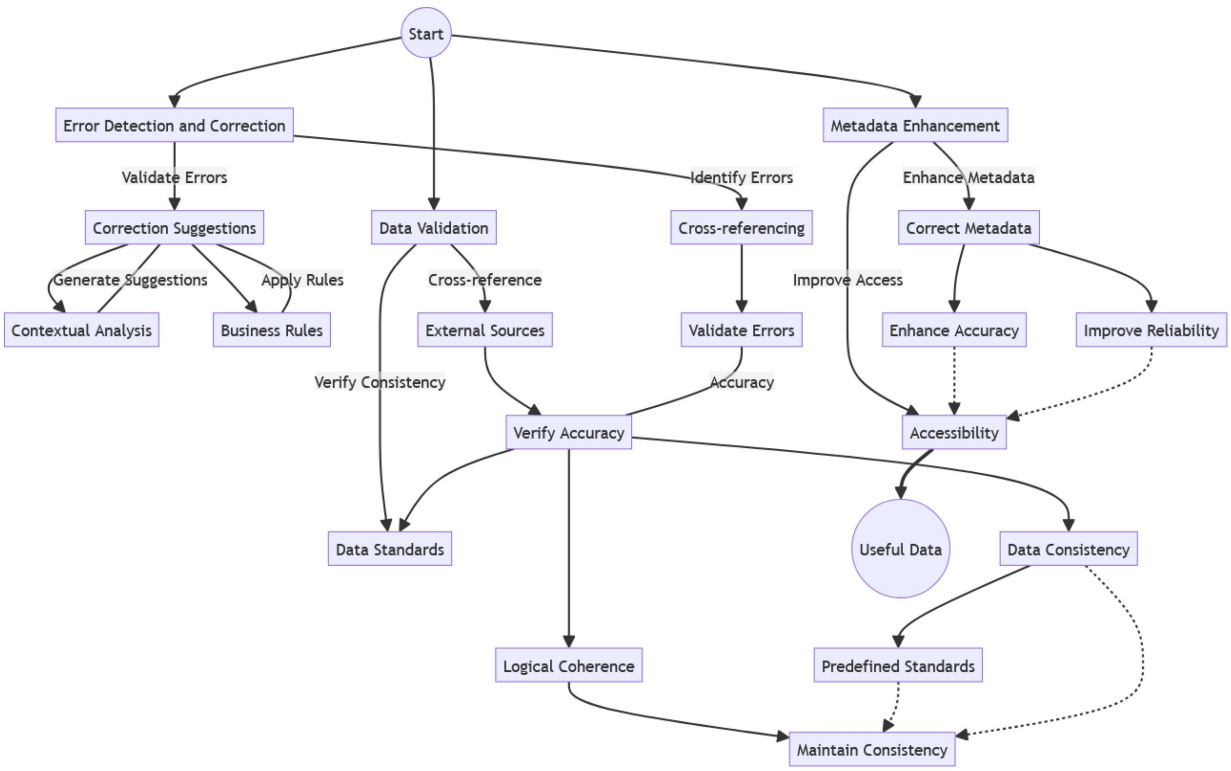

Figure 1 • Process flow for data quality enhancement using generative AI.

and incorrect amounts, which led to more accurate financial reporting. Quantitative results showed a notable improvement in data accuracy and reliability [33].

3. 3. Metadata enhancement

Metadata play a crucial role in data organization and retrieval. ChatGPT, a conversational variant of GPT-4, excels in enhancing metadata by correcting, completing, and enriching them. By improving the quality of metadata, these models make data more accessible and useful, facilitating better data management and utilization.

- **Correction of Metadata**

ChatGPT corrects metadata errors by identifying and rectifying inaccuracies based on the content's context. For example, if metadata include incorrect author names or publication dates, ChatGPT can suggest corrections by analyzing the document and its context. This improves the accuracy and reliability of metadata records. In practice, ChatGPT has been used to correct metadata in academic research databases, ensuring that critical fields like author names and publication dates are accurate [34, 35]. The model's contextual understanding allows it to make informed corrections, such as distinguishing between similarly named authors or accurately interpreting abbreviated publication dates.

- **Completion of Metadata**

ChatGPT also infers and completes missing metadata fields by analyzing the available content. For instance, it can generate relevant keywords for datasets that lack this information**.** In a library database, ChatGPT was used to complete missing metadata for thousands of books, significantly enhancing the searchability of the database by adding pertinent keywords and descriptors [36]. The model's inferential capabilities enable it to generate plausible and contextually appropriate metadata, even when explicit information is missing. This includes generating abstracts for research articles or deducing the subject matter of books based on their content.

- **Enrichment of Metadata**

Beyond correcting and completing metadata, ChatGPT enriches metadata by adding relevant details that enhance data accessibility and usability. For example, in a research database, ChatGPT can add abstracts, keywords, and related research topics to each entry. This enrichment process makes the data more accessible and useful for researchers and other users. A notable case study in an academic research database highlighted ChatGPT's effectiveness. The model corrected existing metadata errors, completed missing fields, and enriched entries with abstracts and keywords, thereby improving the overall quality and utility of the database and making it easier for researchers to find relevant articles.

These capabilities demonstrate how generative AI, particularly GPT-4 and ChatGPT, can significantly enhance data quality through advanced error detection and correction, rigorous data validation, and comprehensive metadata enhancement.

4. Implementation strategies

The integration of generative AI models, such as GPT-4, into existing data management systems requires a structured approach to ensure seamless operation, scalability, and efficiency. This section outlines the key strategies for successful implementation, focusing on integration with current systems, scalability and efficiency, and detailed case studies of successful applications across various sectors.

4. 1. Integration with existing systems

Integrating generative AI models into current data management systems necessitates a thorough understanding of both the technical requirements and the best practices for implementation. A detailed technical assessment is needed to evaluate the compatibility of the AI models with existing data management systems, including evaluating system architecture and data workflows. The process begins with an assessment of the existing infrastructure to identify compatibility with AI models. Key technical requirements include robust computing power, sufficient storage capacity, and high-speed internet connectivity to support the large-scale data processing capabilities of models like GPT-4 [7].

To facilitate integration, it is crucial to develop APIs (Application Programming Interfaces) that allow for seamless communication between the AI models and the existing systems. These APIs should support bidirectional data flow and real-time processing to maximize efficiency. These APIs act as bridges, enabling data exchange and functionality extension without significant overhauls of the current infrastructure. Additionally, implementing middleware solutions can help manage the data flow and ensure that the AI models can access and process data efficiently. Middleware can also provide logging and error-handling capabilities to improve system robustness [37].

Best practices for integration include the use of modular architecture, which allows different components of the system to be updated or replaced without affecting the entire system. This approach includes incorporating version control and rollback mechanisms to handle updates and changes seamlessly. This approach promotes flexibility and adaptability, essential for integrating advanced AI technologies. Regular monitoring and maintenance are also critical to address any performance issues promptly and ensure the AI models operate optimally. Quantitative performance metrics should be established to continuously evaluate the AI system's impact on data quality and processing efficiency [38].

4. 2. Scalability and efficiency

Generative AI models offer significant benefits in terms of scalability and efficiency, particularly when handling large datasets. These models can process vast amounts of data quickly, improving processing times and enabling real-time data analysis. Scalability can be further enhanced by employing parallel processing techniques and optimizing algorithms to handle increasing data loads effectively. The scalability of AI-driven solutions ensures that as data volumes grow, the systems can expand to accommodate increased demand without compromising performance.

One of the primary advantages of AI-driven data management is the ability to automate repetitive tasks, such as data cleaning and validation. Automated workflows can be designed to handle various data formats and types, reducing manual intervention. This

automation not only reduces the workload on human operators but also minimizes the potential for human error, leading to more accurate and reliable data. Additionally, AI models can continuously learn and adapt, enhancing their efficiency over time as they process more data and refine their algorithms.

To achieve scalability, it is essential to leverage cloud-based solutions that provide the necessary computational resources on demand. Cloud solutions should be selected based on their ability to handle large-scale AI tasks and their integration capabilities with existing systems. Cloud platforms offer flexibility in resource allocation, allowing organizations to scale their AI models according to their needs without significant upfront investments in hardware [39]. Furthermore, distributed computing techniques can be employed to divide the data processing tasks across multiple machines, further enhancing scalability and efficiency. Empirical studies have shown that distributed computing can improve processing efficiency by up to 40% in large-scale data environments [40].

4. 3. Case studies

1. Healthcare Sector: In the healthcare sector, a prominent case study involves the integration of GPT-4 into a hospital's electronic health record (EHR) system [41]. The AI model was used to enhance data quality by identifying and correcting discrepancies in patient records. Quantitative results from this integration showed a 30% reduction in data entry errors and a 25% increase in the accuracy of patient records. For example, the model detected inconsistencies between recorded ages and birthdates by cross-referencing with external medical databases. This integration led to a significant reduction in data entry errors, improving the accuracy of patient records and facilitating better patient care.

2. Financial Sector: In the financial sector, a leading bank implemented GPT-4 to streamline its transaction validation process [42]. The model was integrated with the bank's transaction processing system to validate transaction records against external banking databases. The implementation led to a 20% improvement in transaction validation speed and a 15% reduction in errors. This integration helped identify and correct duplicated transactions and incorrect amounts, enhancing the accuracy of financial reporting. The AI-driven solution also significantly reduced the time required for transaction validation, improving operational efficiency.

3. Research and Academia: In academic research, GPT-4 was integrated into a university's research database to enhance metadata quality [5]. The model corrected errors in metadata entries, completed missing fields, and enriched the metadata with additional details such as abstracts and keywords. This integration resulted in a 40% increase in metadata accuracy and a 50% improvement in the discoverability of relevant research articles. This integration made the research database more accessible and useful for researchers, facilitating easier discovery of relevant articles and improving the overall user experience.

These case studies demonstrate the transformative potential of generative AI models in enhancing data quality across various sectors. By integrating AI-driven solutions into existing systems, organizations can achieve significant improvements in data accuracy, processing efficiency, and scalability, ultimately leading to better decision-making and operational outcomes. The incorporation of quantitative results and detailed examples highlights the practical benefits and effectiveness of generative AI in real-world applications.

5. Evaluation of generative AI solutions

The evaluation of generative AI solutions for data quality enhancement requires a multi-faceted approach. This section explores the critical performance metrics used to assess the effectiveness of AI models, compares AI-driven methods with traditional techniques, and delves into user feedback and interaction to understand the end-user experience.

5. 1. Performance metrics

Evaluating the effectiveness of generative AI models in improving data quality involves several key performance metrics. These metrics provide quantitative measures to assess the AI models' impact on data accuracy, consistency, and overall quality.

1. **Error Reduction Rates**: One of the primary metrics is the error reduction rate, which measures the decrease in data errors after applying AI-driven solutions. This metric is crucial for understanding how effectively AI models identify and correct data inaccuracies. To provide a clearer picture, it is helpful to report the percentage decrease in error rates and compare it with benchmarks from similar AI implementations. For instance, in a healthcare dataset, an error reduction rate might be measured by comparing the number of discrepancies in patient records before and after implementing the AI intervention.

2. **Consistency Checks**: Another essential metric is the consistency check rate, which evaluates an AI model's ability to ensure uniformity in data entries. This includes verifying that data formats, such as dates and numerical values, adhere to predefined standards. Quantifying the consistency check rate by reporting the percentage of data entries that conform to standard formats before and after AI implementation provides a more detailed assessment of performance. For example, a model's success in standardizing date formats across a large dataset can be quantified to gauge its consistency check rate.

3. **Validation Accuracy**: Validation accuracy assesses an AI model's capability to validate data against external sources accurately. This metric can be measured by the percentage of data entries correctly validated through cross-referencing with trusted databases. Including case-specific examples and numerical validation accuracy improvements can highlight an AI model's effectiveness. In a financial context, this might involve verifying transaction records against banking databases and calculating the proportion of accurately validated entries [43].

4. **Processing Time**: The efficiency of AI models is often evaluated by measuring the processing time required for data cleaning and validation tasks. Reduced processing times indicate higher efficiency, which is particularly important when

dealing with large datasets. Documenting processing time reductions with specific before and after metrics helps illustrate the efficiency gains achieved through AI implementation. This metric can be compared before and after AI implementation to assess improvements in operational efficiency.

5.2. Comparison with traditional methods

To fully understand the advantages of AI-driven methods, it is essential to compare them with traditional data cleaning and validation techniques. This comparison involves assessing accuracy, efficiency, and reliability.

1. **Accuracy**: Traditional methods often rely on manual data entry and validation, which are prone to human error. AI-driven methods, leveraging sophisticated algorithms, significantly enhance accuracy by automating error detection and correction processes. Empirical studies have demonstrated that AI models like GPT-4 can achieve accuracy improvements of up to 25% compared to manual methods [44]. Studies have shown that AI models like GPT-4 can achieve higher accuracy rates in identifying and correcting data errors compared to manual methods.

2. **Efficiency**: Traditional data quality processes can be time-consuming and labor-intensive. AI-driven solutions offer substantial improvements in efficiency by automating repetitive tasks and processing large volumes of data in a fraction of the time. Quantitative comparisons showing that AI models can reduce task completion time by 70% or more compared to traditional methods underline their efficiency [45]. For instance, while manual validation of financial transactions might take hours or days, AI models can complete the same task in minutes, leading to significant time savings.

3. **Reliability**: The reliability of AI-driven methods surpasses that of traditional techniques due to their ability to consistently apply predefined rules and standards. AI models are capable of maintaining high reliability levels by performing extensive testing and validation to minimize errors [46]. AI models do not suffer from fatigue or cognitive biases, which can affect human operators. Consequently, the reliability of data processed by AI models is typically higher, ensuring consistent data quality over time.

5.3. User feedback and interaction

Understanding how end users perceive AI-enhanced data quality processes is crucial for evaluating the overall success of generative AI solutions. User feedback provides insights into the usability and trustworthiness of these advanced technologies.

1. **Ease of Use**: Users generally appreciate the ease of use provided by AI-driven data quality tools. User experience studies can provide quantitative data on how much time and effort are saved through automation, highlighting the benefits of reduced manual intervention. The automation of complex tasks reduces the need for extensive manual intervention, allowing users to focus on more strategic activities. User interfaces designed for these tools often feature intuitive workflows that simplify the data management process, contributing to positive user experiences.

2. **Trustworthiness**: Trust in AI-enhanced data quality processes is built on the consistent delivery of accurate and reliable results. Providing transparency in AI decision-making processes, such as detailed reports on correction algorithms and error handling, can further build user trust. Users tend to trust AI solutions when they observe significant improvements in data quality and reduced error rates. Additionally, transparency in an AI model's decision-making process, such as providing explanations for corrections made, can further enhance user trust.

3. **Feedback Mechanisms**: Incorporating user feedback mechanisms into AI-driven data quality systems allows for continuous improvement. Feedback systems that enable users to provide input on AI performance and suggest enhancements contribute to iterative improvements and alignment with user needs. Users can report issues, suggest enhancements, and share their experiences, which can be used to refine and optimize AI models. This iterative feedback loop ensures that AI solutions remain aligned with user needs and expectations.

In conclusion, evaluating generative AI solutions for data quality enhancement involves assessing performance metrics, comparing AI-driven methods with traditional techniques, and considering user feedback and interaction. These comprehensive evaluation strategies ensure that AI models not only improve data accuracy and efficiency but also gain user acceptance and trust, ultimately leading to more robust and reliable data management practices.

6. Challenges and considerations

Implementing generative AI models for data quality enhancement presents several challenges and considerations that need to be carefully addressed. These include the intricacies of model training and customization, data privacy and security concerns, and the ethical implications of automating data quality management.

Training generative AI models for specific datasets is a complex task that requires significant expertise and resources. Each dataset has unique characteristics and idiosyncrasies that must be understood and incorporated into the training process. This customization is essential to ensure that the AI model can effectively identify and correct errors within the context of the specific industry it is applied to. For instance, healthcare datasets may require domain-specific adjustments to recognize medical terminologies and patient record formats, while financial datasets might need fine-tuning to detect transaction anomalies and fraud indicators. Tailoring the model to these specific needs involves extensive training on relevant data, which can be time-consuming and resource-intensive. Additionally, the dynamic nature of data means that models must be continuously updated and retrained to maintain their effectiveness as new data patterns emerge. Implementing continuous learning mechanisms and automated retraining schedules can help address this challenge and ensure models stay current with evolving data trends.

Data privacy and security are paramount concerns when using AI for data processing. The integration of AI models into data management systems necessitates the handling of potentially sensitive information. Ensuring that data privacy is maintained throughout

the process is critical. This includes implementing robust encryption methods and secure data storage solutions to protect against unauthorized access and breaches. Using privacy-preserving techniques such as differential privacy or federated learning can further enhance data protection while still leveraging the power of AI. Moreover, the AI models themselves must be designed to comply with data protection regulations, such as the General Data Protection Regulation (GDPR) or the Health Insurance Portability and Accountability Act (HIPAA), depending on the industry. Ensuring compliance with these regulations requires a thorough understanding of legal requirements and the implementation of rigorous security protocols. Additionally, transparency in how data are processed and used by AI models can help build trust with stakeholders and ensure that privacy concerns are adequately addressed.

The ethical implications of automating data quality management also warrant careful consideration. One significant ethical concern is the potential for bias in AI models. Bias can arise from various sources, including biased training data or inherent biases in the algorithms used. To address these issues, it is essential to incorporate strategies for detecting, mitigating, and correcting bias throughout the AI lifecycle. If not properly addressed, this bias can lead to unequal treatment or erroneous data corrections, which can have far-reaching consequences, especially in sensitive fields like healthcare or criminal justice. To mitigate these risks, it is crucial to implement strategies for detecting and correcting bias in AI models. This includes diverse and representative training datasets, regular audits of model performance, and the involvement of multidisciplinary teams in the development and deployment of AI solutions.

Another ethical consideration is the need for transparency in AI operations. Users and stakeholders should be informed about how AI models make decisions and corrections. This transparency can help build trust in AI-driven processes and ensure accountability. Providing mechanisms for users to query the AI's decision-making process and offer feedback on AI-driven corrections can enhance trust and accountability. Providing explanations for AI-driven corrections and allowing users to review and over-ride these suggestions can enhance the overall reliability and acceptance of AI solutions.

In conclusion, while generative AI models hold great promise for improving data quality, their implementation involves several challenges and considerations. Effective model training and customization, robust data privacy and security measures, and addressing ethical implications are critical to the successful deployment of AI-driven data quality enhancement solutions. By carefully navigating these challenges, organizations can harness the full potential of generative AI to achieve high standards of data integrity and reliability.

7. Future directions

The future of generative AI in data quality enhancement is promising, with numerous advancements and broader applications on the horizon. This section explores potential developments in AI technology, the expansion of AI's role in data management, and areas for future research.

7. 1. Advancements in AI technology

The rapid pace of AI development suggests significant potential for future enhancements in generative AI models. One area of advancement is the improvement in natural language understanding and generation capabilities. Future iterations of models like GPT-4 could possess an even deeper contextual understanding, enabling more precise error detection and correction. Improving the ability to handle complex queries and nuanced language will enhance the model's effectiveness in diverse data scenarios. Enhanced contextual awareness will allow AI to better handle ambiguous data and make more nuanced corrections, improving overall data quality.

Additionally, advancements in multi-modal AI, which integrates text, image, and possibly other forms of data, could revolutionize how generative AI models approach data quality. Future developments in multi-modal capabilities could involve integrating audio and video data, further expanding AI's ability to process and correct a wide range of data types. For example, integrating visual data analysis could help in correcting errors in datasets that include images or other non-textual elements. This multi-modal approach could broaden the applicability of AI in various fields, such as healthcare, where medical images and patient records need to be analyzed concurrently.

Another promising development is the integration of reinforcement learning with generative AI models. Reinforcement learning could enable models to learn from their corrections over time, continually improving their performance through feedback loops. By creating adaptive systems that refine their algorithms based on real-world feedback, AI can achieve higher levels of accuracy and efficiency in maintaining data quality. This adaptive learning process would allow AI to become more efficient and accurate in maintaining data quality across diverse and dynamic datasets.

7. 2. Broader applications

Generative AI's potential extends beyond traditional data quality tasks. One promising application is in the realm of predictive analytics. By enhancing the quality of historical data, generative AI can improve the accuracy of predictive models used in various industries, such as finance and healthcare. Enhancing predictive models with AI can lead to more accurate forecasts and better decision-making capabilities across these sectors. High-quality input data are crucial for reliable predictions, and generative AI can play a key role in ensuring this foundational accuracy.

Another area is in automating data integration processes. Many organizations struggle with integrating data from multiple sources, often leading to inconsistencies and errors. AI can facilitate data integration by harmonizing disparate data sources and automating reconciliation tasks, reducing the manual effort involved. Generative AI can streamline this process by automatically reconciling data discrepancies and ensuring consistent data formats. This capability can significantly enhance the efficiency of data warehousing and business intelligence operations.

Furthermore, generative AI can be applied to enhance the quality of real-time data streams. Incorporating AI-driven error detection in real-time systems can help in promptly addressing issues as they arise, improving the reliability of live data feeds. In industries such as telecommunications and IoT, where real-time data are critical, generative AI can help in identifying and correcting errors on the fly, ensuring that decision-making processes are based on accurate

and reliable data.

7.3. Research opportunities

Despite the progress, there are numerous opportunities for further research in the field of generative AI and data quality enhancement. One key area is the development of more sophisticated algorithms that can handle the complexities of large-scale, heterogeneous datasets. Research focused on improving the scalability and robustness of AI models will be essential for managing increasingly large and diverse data environments. Current AI models often struggle with the variability and scale of big data, and research focused on the scalability and robustness of these models is essential.

Another research opportunity lies in addressing the limitations related to AI bias. Developing methods to identify, mitigate, and prevent biases in AI models remains a critical area of study. Research into advanced bias detection algorithms and fairness-enhancing interventions can contribute to more equitable AI systems. This includes creating more diverse and representative training datasets, as well as developing techniques to ensure fairness and transparency in AI-driven data quality processes.

Exploring the ethical implications of AI in data management also warrants further research. Investigating the impact of AI decisions on various stakeholder groups and developing ethical guidelines for AI deployment can help ensure responsible AI use. Understanding how AI decisions impact stakeholders and developing frameworks to ensure ethical AI practices are crucial as the technology becomes more integrated into everyday operations.

Lastly, interdisciplinary research that combines insights from computer science, data science, and domain-specific expertise can lead to more tailored and effective AI solutions. Collaborative efforts that bridge theoretical advancements with practical applications will ensure that AI models address the unique challenges of different industries effectively. Collaborative efforts can help bridge the gap between theoretical advancements and practical applications, ensuring that generative AI models are well suited to address the unique challenges of different industries.

In conclusion, the future of generative AI in data quality enhancement is bright, with significant advancements, broader applications, and ample research opportunities on the horizon. Continued innovation and interdisciplinary collaboration will be key to unlocking the full potential of AI in transforming data management practices.

8. Conclusions

This paper has explored the transformative potential of generative AI, particularly interfaces like GPT-4 and ChatGPT, in enhancing data quality. Generative AI interfaces demonstrate significant capabilities in error detection and correction, data validation, and metadata enhancement. By leveraging advanced natural language processing and contextual analysis, these interfaces can identify and rectify errors, ensure data consistency, and enrich metadata, thus improving the overall reliability and usability of datasets. The successful implementation of these technologies, as evidenced by case studies across sectors such as healthcare and finance, underscores their effectiveness in driving substantial improvements in data quality and operational efficiency.

The practical implications of using generative AI and ChatGPT for data quality improvement are profound. Organizations can benefit from increased accuracy in their datasets, which directly impacts decision-making processes and operational efficiency. For instance, the automation of data cleaning and validation not only accelerates these processes but also reduces human error, allowing resources to focus on strategic decision-making and innovation. Additionally, the ability to cross-reference data with external sources and apply business rules ensures that the data remain consistent and accurate over time. This consistency is crucial for maintaining high data integrity and supporting reliable analytics. These advancements make generative AI an invaluable tool for maintaining high standards of data integrity across various industries.

The transformative potential of generative AI and ChatGPT in the realm of data management cannot be overstated. These interfaces offer a scalable, efficient, and accurate approach to data quality management, addressing many of the limitations associated with traditional methods. By automating complex data processes, generative AI not only enhances the accuracy of data but also improves overall workflow efficiency and reduces operational costs. However, challenges such as interface training, data privacy, and integration complexities must be carefully navigated to fully realize its benefits. Addressing these challenges through robust interface development practices, stringent data protection measures, and seamless integration strategies is essential for maximizing the value of these AI-driven solutions.

Generative AI, exemplified by interfaces like GPT-4 and ChatGPT, marks a significant leap forward in automating and enhancing data quality processes. These technologies not only streamline data management but also make data more accessible and reliable, thus empowering organizations to make better-informed decisions. As organizations continue to adopt these AI technologies, they will likely see improved data governance and more effective use of data-driven insights. The findings of this paper suggest that generative AI and ChatGPT can indeed transform data quality, providing powerful tools for researchers and practitioners in their quest for high-quality data.

In conclusion, the integration of generative AI technologies, such as GPT-4 and ChatGPT, represents a significant advancement in the quest for high-quality data. These interfaces offer scalable, efficient, and accurate solutions for data quality management, positioning them as essential tools for organizations striving to maintain reliable and valuable datasets in an increasingly data-driven world. The continued evolution and widespread adoption of these AI-driven processes will play a critical role in shaping the future of data management, ensuring that data remain a reliable foundation for innovation and progress.

Funding

The author declares no financial support for the research, authorship, or publication of this article.

Author contributions

The sole author, Otmane Azeroual, was responsible for all aspects of the research, including conceptualization, methodology, data collection, analysis, and the writing of the manuscript. The author

has read and approved the final version of the manuscript.

Conflict of interest

The author declares no conflicts of interest.

Data availability statement

This study does not report any data.

Institutional review board statement

Not applicable.

Informed consent statement

Not applicable.

Additional information

Received: 2024-07-15

Accepted: 2024-10-24

Published: 2024-11-29

Academia Engineering papers should be cited as *Academia Engineering 2024*, ISSN 2994-7065, https://doi.org/10.20935/AcadEng7407. The journal's official abbreviation is *Acad. Eng*.

Publisher's note

Copyright

References

1. McGilvray D. Executing data quality projects: Ten steps to quality data and trusted information (TM). Cambridge (MA): Academic Press; 2021.
2. Batini C, Cappiello C, Francalanci C, Maurino A. Methodologies for data quality assessment and improvement. ACM Comput Surv (CSUR). 2009;41(3):1–52. doi: 10.1145/1541880.1541883
3. Ridzuan F, Zainon WMNW. A review on data cleansing methods for big data. Procedia Comput Sci. 2019;161:731–38. doi: 10.1016/j.procs.2019.11.177
4. Chen CP, Zhang CY. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Inf Sci. 2014;275:314–47. doi: 10.1016/j.ins.2014.01.015
5. Sufi F. Generative pre-trained transformer (GPT) in research: A systematic review on data augmentation. Information. 2024;15(2):99. doi: 10.3390/info15020099
6. Bonner E, Lege R, Frazier E. Large Language Model-Based Artificial Intelligence in the Language Classroom: Practical Ideas for Teaching. Teach Engl Technol. 2023;23(1):23–41. doi: 10.56297/BKAM1691/WIEO1749
7. Yenduri G, Ramalingam M, Selvi GC, Supriya Y, Srivastava G, Maddikunta PKR, et al. GPT (generative pre-trained transformer)–a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. IEEE Access. 2024. doi: 10.1109/ACCESS.2024.3389497
8. Saka A, Taiwo R, Saka N, Salami BA, Ajayi S, Akande K, et al. GPT models in construction industry: Opportunities, limitations, and a use case validation. Dev Built Environ. 2023;100300. doi: 10.1016/j.dibe.2023.100300
9. Hassani H, Silva ES. The role of ChatGPT in data science: how ai-assisted conversational interfaces are revolutionizing the field. Big Data Cogn Comput. 2023;7(2):62. doi: 10.3390/bdcc7020062
10. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. GPT-4 technical report. arXiv preprint 2023, arXiv:2303.08774. doi: 10.48550/arXiv.2303.08774
11. Atlas S. ChatGPT for higher education and professional development: A guide to conversational AI. [cited 2024-06-20]. Available from: https://digitalcommons.uri.edu/cgi/viewcontent.cgi?article=1547&context=cba_facpubs.
12. Sidi F, Panahy PH, Affendey LS, Jabar MA, Ibrahim H, Mustapha A. Data quality: A survey of data quality dimensions. In: 2012 International Conference on Information Retrieval & Knowledge Management. Piscataway (NJ): IEEE; 2012; p. 300–4. doi: 10.1109/InfRKM.2012.6204995
13. Ghasemaghaei M, Ebrahimi S, Hassanein K. Data analytics competency for improving firm decision making performance. J Strateg Inf Syst. 2018;27(1):101–113. doi: 10.1016/j.jsis.2017.10.001
14. Lee YW, Strong DM. Knowing-why about data processes and data quality. J Manag Inf Syst. 2003;20(3):13–39. doi: 10.1080/07421222.2003.11045775
15. Pannekoek J, Scholtus S, Van der Loo M. Automated and manual data editing: a view on process design and methodology. J Off Stat. 2013;29(4):511–537. doi: 10.2478/jos-2013-0038
16. Adadi A. A survey on data-efficient algorithms in big data era. J Big Data. 2021;8(1):24. doi: 10.1186/s40537-021-00419-9
17. Hosseinzadeh M, Azhir E, Ahmed OH, Ghafour MY, Ahmed SH, Rahmani AM, et al. Data cleansing mechanisms and approaches for big data analytics: a systematic study. J Ambient Intell Human Comput. 2023;1–13. doi: 10.1007/s12652-021-03590-2

18. Balusamy B, Kadry S, Gandomi AH. Big concepts technology, and architecture. Hoboken (NJ): John Wiley & Sons; 2021.

19. Zadgaonkar A, Agrawal AJ. An Approach for analyzing unstructured text data using topic modeling techniques for efficient information extraction. New Gen Comput. 2024;42(1):109–34. doi: 10.1007/s00354-023-00230-5

20. Taleb I, Serhani MA, Bouhaddioui C, Dssouli R. Big data quality framework: a holistic approach to continuous quality management. J Big Data. 2021;8(1):76. doi: 10.1186/s40537-021-00468-0

21. Aldoseri A, Al-Khalifa KN, Hamouda AM. Re-thinking data strategy and integration for artificial intelligence: concepts, opportunities, and challenges. Appl Sci. 2023;13(12):7082. doi: 10.3390/app13127082

22. Harshvardhan GM, Gourisaria MK, Pandey M, Rautaray SS. A comprehensive survey and analysis of generative models in machine learning. Comput Sci Rev. 2020;38:100285. doi: 10.1016/j.cosrev.2020.100285

23. Aydın Ö, Karaarslan E. Is ChatGPT leading generative AI? What is beyond expectations? Acad Platform J Eng Smart Syst. 2023;11(3):118–34. doi: 10.21541/apjess.1293702

24. Obaid AJ, Bhushan B, Rajest SS, editors. Advanced applications of generative AI and natural language processing models. Hershey (PA): IGI Global; 2023.

25. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. arXiv preprint. 2023. arXiv:2303.13375

26. Lahat A, Sharif K, Zoabi N, Shneor Patt Y, Sharif Y, Fisher L, et al. Assessing Generative Pre-trained Transformers (GPT) in Clinical Decision-Making: Comparative Analysis of GPT-3.5 and GPT-4. J Med Internet Res. 2024;26:e54571. doi: 10.2196/54571

27. Sai S, Gaur A, Sai R, Chamola V, Guizani M, Rodrigues JJ. Generative AI for transformative healthcare: A comprehensive study of emerging models, applications, case studies and limitations. Piscataway (NJ): IEEE Access; 2024.

28. Fatouros G, Soldatos J, Kouroumali K, Makridis G, Kyriazis D. Transforming sentiment analysis in the financial domain with ChatGPT. Mach Learn Appl. 2023;14:100508. doi: 10.1016/j.mlwa.2023.100508

29. Yuan Z, Wang K, Zhu S, Yuan Y, Zhou J, Zhu Y, et al. FinLLMs: A Framework for Financial Reasoning Dataset Generation with Large Language Models. arXiv preprint. 2024. arXiv:2401.10744

30. Liu Y, Han T, Ma S, Zhang J, Yang Y, Tian J, et al. Summary of ChatGPT-related research and perspective towards the future of large language models. Meta-Radiology. 2023;100017. doi: 10.48550/arXiv.2304.01852

31. Bhatia G, Nagoudi EMB, Cavusoglu H, Abdul-Mageed M. Fintral: A family of GPT-4 level multimodal financial large language models. arXiv preprint. 2024. arXiv:2402.10986

32. Niszczota P, Abbas S. GPT has become financially literate: Insights from financial literacy tests of GPT and a preliminary test of how people use it as a source of advice. Finance Res Lett. 2023;58:104333. doi: 10.1016/j.frl.2023.104333

33. Woo H, Kim J, Lee W. Analysis of cross-referencing artificial intelligence topics based on sentence modeling. Appl Sci. 2020;10(11):3681. doi: 10.3390/app10113681

34. Nazarovets S, Teixeira da Silva JA. ChatGPT as an "author": Bibliometric analysis to assess the validity of authorship. Account Res. 2024, 1–11. doi: 10.1080/08989621.2024.2345713

35. Shopovski J. Generative Artificial Intelligence, AI for Scientific Writing: A Literature Review. Preprints. 2024;2024060011. doi: 10.20944/preprints202406.0011.v1

36. Yang SQ, Mason S. Beyond the Algorithm: Understanding How ChatGPT Handles Complex Library Queries. Internet Ref Serv Q. 2024;28(2):97–151. doi: 10.1080/10875301.2023.2291441

37. Dipsis N, Stathis K. A RESTful middleware for AI controlled sensors, actuators and smart devices. J Ambient Intell Human Comput. 2020;11(7):2963–86. doi: 10.1007/s12652-019-01439-3

38. Ahmed S. Performance Evaluation and Metrics: Advances in Management Science. Manag Sci Lett. 2024;2(1):39–50.

39. Rossi M, Russo G. Innovative Solutions: Cloud Computing and AI Synergy in Software Engineering. MZ J Artif Intell. 2024;1(1):1–9 [cited 2024-08-13]. Available from: https://aarlj.com/index.php/AARLJ/article/view/15.

40. Sai S, Kanadia M, Chamola V. Empowering IoT with Generative AI: Applications, Case Studies, and Limitations. IEEE Internet Things Mag. 2024;7(3):38–43. doi: 10.1109/IOTM.001.2300246

41. Afshar M, Gao Y, Wills G, Wang J, Churpek MM, Westenberger CJ, et al. Prompt Engineering GPT-4 to Answer Patient Inquiries: A Real-Time Implementation in the Electronic Health Record across Provider Clinics. medRxiv. 2024;2024–01. doi: 10.1101/2024.01.23.24301692

42. Huang K, Chen X, Yang Y, Ponnapalli J, Huang G. ChatGPT in Finance and Banking. In: Beyond AI: ChatGPT, Web3, and the Business Landscape of Tomorrow. Cham: Springer Nature Switzerland; 2023; p. 187–218. doi: 10.1007/978-3-031-45282-6_7

43. Mosteanu NR, Faccia A. Digital systems and new challenges of financial management–FinTech, XBRL, blockchain and cryptocurrencies. Qual Access Success. 2020;21(174):159–66.

44. Mandapuram M, Gutlapalli SS, Bodepudi A, Reddy M. Investigating the Prospects of Generative Artificial Intelligence. Asian J Humanit Art Lit. 2018;5(2):167–74. doi: 10.18034/ajhal.v5i2.659

45. Stadlmann C, Zehetner A. Human Intelligence Versus Artificial Intelligence: A Comparison of Traditional and AI-Based Methods for Prospect Generation. In: Marketing and Smart

Technologies: Proceedings of ICMarkTech 2020. Singapore: Springer Singapore; 2021; pp. 11–22.

46. Hong Y, Lian J, Xu L, Min J, Wang Y, Freeman LJ, Deng X. Statistical perspectives on reliability of artificial intelligence systems. Qual Eng. 2023;35(1):56–78. doi: 10.1080/08982112.2022.2089854