AI-Powered Training Data Curation Bot

Mr. R. Vikram[1], Jeenokanth G[2], Mohammed Unais N[3], Lohit U[4]
[1]*M.E, Assistant Professor, Adhiyamaan College of Engineering (Autonomous), Hosur*
[2,3,4]*UG Students, Adhiyamaan College of Engineering (Autonomous), Hosur*

Abstract*—The Training Data Curation Bot is a comprehensive and intelligent AI-driven system designed to automate the process of generating high-quality training datasets from unstructured documents such as PDFs, text files, web pages, and spreadsheets. The primary objective of the system is to simplify and streamline the preparation of training data required for fine-tuning and building domain-specific AI models, eliminating the need for manual data extraction and formatting. The system provides users with an intuitive interface to upload documents through a centralized platform. Once uploaded, the documents are automatically processed using a robust document loading pipeline that detects file types, extracts textual content, and converts it into a standardized internal format. The extracted content is further cleaned, segmented into meaningful text chunks, and prepared for AI driven task execution. Using predefined templates and specialized AI task generators, the system automatically creates structured training examples such as question–answer pairs, summaries, and classification data. Each generated training example undergoes a quality evaluation process to ensure relevance, consistency, and usability. The validated data is then organized into datasets with configurable training, validation, and testing splits, making it directly usable for machine learning and large language model (LLM) fine-tuning. The application supports both command-line and web-based interaction, featuring a modern dashboard that enables real-time monitoring of document processing, training data generation, and quality metrics. The backend is implemented using Python and asynchronous processing techniques to ensure high performance, scalability, and efficient resource management, while the frontend dashboard provides a user-friendly and visually intuitive experience. This automated system significantly reduces human effort, minimizes errors, and improves the speed and reliability of training data preparation. It ensures transparency, traceability, and scalability by maintaining structured data models, detailed logs, and comprehensive quality reports. By transforming raw documents into high-quality AI-ready datasets, the Training Data Curation Bot demonstrates the practical application of artificial intelligence, natural language processing, and full-stack system design.

I. INTRODUCTION

In today's data-driven and artificial intelligence–oriented era, automation plays a vital role in simplifying complex data preparation processes required for building intelligent systems. One of the major challenges in developing AI and machine learning models is the manual creation of high-quality training datasets from unstructured sources such as documents, PDFs, web pages, and text files. This traditional approach is time-consuming, error-prone, and requires significant human effort to extract, clean, and structure data. To address these challenges, the Training Data Curation Bot has been developed as an automated and intelligent system that streamlines the entire process of transforming raw documents into AI-ready training datasets. The system provides users with a user-friendly interface to upload various types of documents through a centralized platform. Once uploaded, the documents are automatically processed using a document loading pipeline that identifies file formats, extracts textual content, and converts it into a standardized structure. The extracted text is further cleaned and divided into meaningful chunks, which are then processed by specialized AI task generators to produce structured training examples such as question answer pairs, summaries, and classification data. These generated outputs are organized into datasets suitable for machine learning and large language model fine-tuning. A key feature of this system is its built-in quality evaluation mechanism, which assesses the relevance, consistency, and usability of each generated training example before it is included in the final dataset. The system also provides real-time monitoring and visual feedback through a web-based dashboard, enabling users to track document processing status, dataset generation progress, and quality metrics. This ensures transparency, traceability, and efficient control over the entire data

generation workflow. 1 The Training Data Curation Bot is implemented using Python for backend processing and AI integration, ensuring efficient asynchronous execution, scalability, and secure handling of data. The frontend dashboard offers a clean and responsive interface that allows users to interact with the system effortlessly. By automating the training data preparation process, the system significantly reduces manual effort, minimizes errors, and accelerates AI development workflows while promoting consistency and reusability of data. The automation also improves overall system reliability by maintaining uniform data structures and reducing dependency on manual intervention. This results in faster turnaround times and improved accuracy in dataset creation. The system architecture follows a modular and scalable design, allowing each component—document loading, text preprocessing, task execution, quality evaluation, and dataset export—to operate independently. Role-based access and structured data models ensure controlled usage, secure data handling, and easy system maintenance. This design also supports seamless integration of new features without affecting existing functionalities. Such modularity enhances system maintainability and allows developers to isolate and resolve issues efficiently. The future scope of this project extends beyond basic training data generation. With further enhancements, the system can support multi-language document processing, advanced analytics to monitor data quality trends, and integration with enterprise AI platforms and cloud-based storage solutions. Additionally, incorporating real-time notifications, collaborative workflows, and deployment-ready export formats can further enhance usability. By automating a critical step in AI development, the Training Data Curation Bot contributes to efficient, transparent, and scalable artificial intelligence solutions while supporting the growing demand for intelligent, data-driven applications. These enhancements will further strengthen the system's adaptability and long-term relevance in evolving AI ecosystems.

II. LITERATURE SURVEY

1. WEB-BASED TRAINING DATA CURATION SYSTEM.

In today's artificial intelligence–driven environment, preparing high-quality training data manually consumes significant time and effort for developers and data [7] engineers. Traditional dataset creation methods require extracting information from unstructured documents, cleaning data, and formatting it manually, often leading to inconsistencies, errors, and delays. To address these challenges, a web-based training data curation system has been proposed. This system allows users to upload documents through an online interface, which are then processed automatically to extract, clean, and structure data. The system maintains organized datasets, supports real-time processing status, and enables easy export of training-ready data. By automating the entire workflow, the system reduces manual workload, improves data quality, and enhances efficiency in AI development processes.

2. ROLE-BASED DIGITAL PLATFORM FOR DATA CURATION MANAGEMENT.

AI data preparation [4] requires structured workflows to ensure accuracy, security, and accountability. A role-based digital [1] platform provides a secure and organized approach to managing training data curation activities based on defined user roles such as administrators and data engineers. Users can upload documents, configure processing parameters, and monitor dataset generation through controlled access. Administrative users can manage system settings and oversee data processing activities, ensuring compliance with data handling policies. Role-based access control enhances system [11] security and prevents unauthorized usage. This structured approach improves reliability, accountability, and efficient management [6] of AI training data workflows.

3. ONLINE PORTAL FOR DOCUMENT-BASED DATA PROCESSING.

Processing unstructured documents for AI training purposes is often complex and resource intensive. The online portal for document-based data processing provides a centralized digital solution where users can upload various document [8] formats through a secure [15] web interface. The system automatically extracts textual content, preprocesses the data, and converts it into structured formats suitable for training [12] datasets. Generated outputs such as question–answer pairs and summaries are stored and made available for

download. The portal also allows users to monitor processing [18] progress in real time, ensuring transparency and reducing manual intervention. This digital approach streamlines document processing and supports efficient dataset creation.

4. SMART AI-DRIVEN DATA GENERATION AND EVALUATION SYSTEM.

Manual generation of training examples often lacks consistency and scalability. The smart AI-driven data generation [10] and evaluation system uses automated pipelines [16] to generate structured training data from processed documents. The system applies AI task generators to create meaningful datasets and evaluates the quality of generated outputs before final storage. Quality checks ensure relevance, accuracy, and usability of the data. By integrating intelligent19] evaluation mechanisms, the system improves dataset [17] reliability while minimizing human errors. This smart automation enhances overall data quality and supports scalable AI model [14] development.

5. AUTOMATED WORKFLOW SYSTEM FOR DATASET CREATION.

Traditional dataset creation workflows are time-consuming and prone to inefficiencies. The automated workflow system for dataset creation provides a stream [2] in ed process where document ingestion, preprocessing, AI task execution, and dataset export are handled sequentially without manual intervention. The system maintains complete records of processed documents and generated datasets, allowing users to track workflow [9] status in real time. Automation reduces processing delays, ensures consistency, and improves productivity. By eliminating repetitive manual tasks, the system enables developers to focus on model development and experimentation rather than data preparation.

6. DIGITAL TRANSFORMATION IN AI TRAINING DATA MANAGEMENT.

With the growing demand for intelligent applications, digital transformation in AI [3] training data management has become essential. This project replaces manual data preparation methods with an automated, web-based [20] platform that integrates document processing [13], AI-driven data generation, and secure data storage. Features such as role-based access, real time tracking, and structured dataset management ensure transparency and efficiency. By adopting a digital transformation approach, organizations can improve data reliability, accelerate AI development cycles, and support scalable AI solutions. The system also lays a foundation for future enhancements such as analytics, cloud [5] integration, and collaborative workflows.

III. SYSTEM ANALYSIS

EXISTING SYSTEM Before the development of the Training Data Curation Bot, most organizations and developers relied on manual and semi-automated methods to prepare training datasets for artificial intelligence and machine learning models. In the existing approach, data engineers had to collect unstructured data from multiple sources such as documents, PDFs, and text files, manually extract relevant information, clean the data, and format it into usable datasets. This process was highly time-consuming and prone to inconsistencies, errors, and data redundancy. Tracking the progress of dataset preparation was difficult, as there was no centralized system to monitor document processing or data generation stages. Developers often had to reprocess documents multiple times due to incorrect formatting or missing data, leading to increased effort and delays in AI model development.

The absence of an automated and structured data curation system also made it challenging to maintain version control and historical records of generated datasets. Organizations faced difficulties in ensuring data quality, consistency, and traceability of training samples back to their original sources. Moreover, the lack of standardized workflows and quality evaluation mechanisms resulted in unreliable or biased datasets. Existing methods also raised concerns regarding data security and controlled access, as sensitive documents were often handled manually without proper validation or protection. Overall, the traditional approach to training data preparation was inefficient, error-prone, and not scalable, highlighting the need for an automated, secure, and intelligent training data curation system.

IV. DISADVANTAGES OF EXISTING SYSTEM

The training data preparation process is time-consuming and heavily dependent on manual data extraction and formatting, leading to significant delays.

There is no centralized system to track document processing stages, making it difficult to monitor dataset generation progress. Manual handling of unstructured data increases the risk of inconsistencies, data loss, and poor data quality across datasets. Human errors such as incorrect labeling, incomplete data extraction, or improper formatting often occur during manual dataset creation. The absence of standardized workflows and quality evaluation mechanisms reduces transparency and reliability in training data generation. Limited data security and access control make sensitive documents vulnerable to unauthorized access or misuse.

V. PROPOSED SYSTEM

The proposed Training Data Curation Bot is a comprehensive and fully automated platform designed to address the limitations of traditional manual training data preparation methods. It streamlines and automates the entire data curation workflow, from document upload to the generation of structured, AI-ready training datasets, through a web-based interface. Users can upload various unstructured documents such as PDFs and text files by providing necessary configuration details through the portal. Once the documents are submitted, the system automatically performs text extraction, cleaning, and preprocessing operations. The processed content is then passed to AI task generators, which create structured training examples such as question–answer pairs, summaries, and classification data. These generated outputs are evaluated for quality and relevance before being organized into final datasets ready for use in machine learning and AI model training. Another important feature of the system is its modular and role-based access mechanism, which ensures secure usage and controlled data handling. The frontend dashboard provides a clean, responsive, and user-friendly interface that allows users to monitor processing progress and download curated datasets easily. The backend, developed using Python, manages document processing, AI task execution, quality evaluation, and secure data storage efficiently. This proposed system eliminates repetitive manual effort, reduces errors, and significantly improves productivity in AI development workflows. It offers a transparent and traceable data curation process while supporting scalability and future enhancements. Furthermore, the system promotes efficient and sustainable data management practices by automating a critical stage of the artificial intelligence development lifecycle.

ADVANTAGES OF PROPOSED SYSTEM:

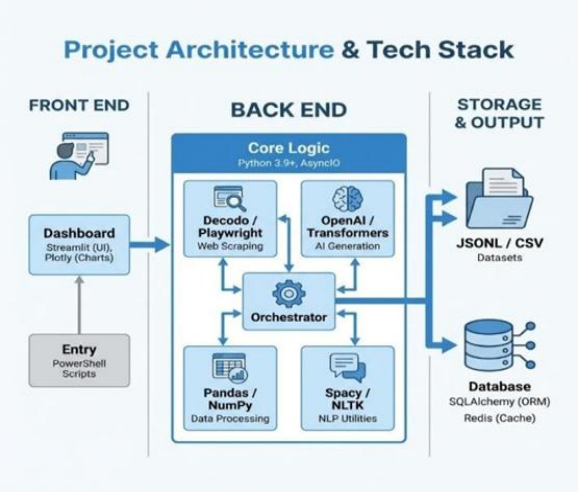
Eliminates the need for manual data extraction and dataset preparation, reducing effort and improving overall efficiency. Automates the entire training data curation workflow, significantly reducing the time required to generate AI-ready datasets. Automatically generates structured and consistent training data from unstructured documents, minimizing errors and inconsistencies. Provides secure, role-based access control, preventing unauthorized usage and ensuring safe handling of sensitive data. Offers a responsive and user-friendly web interface that allows users to monitor processing progress and download curated datasets easily.

VI. PROPOSED SOLUTION

The proposed solution aims to fully automate the training data preparation process through an intuitive, user-centric web application. The system is designed to ensure efficiency, accuracy, and transparency across all stages of data curation. Users can upload unstructured documents and monitor the processing workflow, while the system automatically handles text extraction, preprocessing, and data transformation tasks. Real-time validation mechanisms ensure that uploaded documents meet required formats and processing criteria. Role-based access control allows only authorized users to interact with the system, ensuring secure and controlled usage. The automated pipeline significantly reduces manual intervention and accelerates the generation of AI-ready training datasets. The solution also follows a structured and modular workflow, where each stage of data processing—document ingestion, text preprocessing, AI-driven task execution, and quality evaluation—operates in a defined sequence. The system dashboard displays processing status, generated outputs, and

quality metrics, enabling users to track progress and identify issues easily. From a development perspective, the solution leverages scalable backend technologies implemented in Python to manage asynchronous processing, AI integration, and secure data storage. The frontend provides a clean and responsive interface, allowing users to interact with the system efficiently across different devices. This project not only reduces the effort and time involved in manual training data preparation but also introduces consistency, traceability, and reliability into AI development workflows. It supports digital transformation by replacing fragmented, manual processes with a unified and automated platform. By promoting standardized data generation, secure handling of documents, and transparent processing, the Training Data Curation Bot serves as a robust and scalable solution for modern artificial intelligence applications.

VII. MODULES



VIII. IMPLEMENTATION

DATA COLLECTION

In the context of the Training Data Curation Bot, data collection plays a critical role in transforming unstructured documents into high-quality, AI-ready training datasets. The system collects data from multiple sources, including user uploads, system-generated outputs, and processing metadata, which together support the complete data curation workflow.

User-Provided Data

Users upload unstructured documents such as PDFs, text files, or other supported formats through the web interface. Along with the documents, users may provide configuration inputs such as dataset type, processing parameters, and task preferences. This user-provided data serves as the primary input for the document processing and training data generation pipeline.

System-Generated Data

The system automatically generates intermediate and final data during processing. This includes extracted text, cleaned and preprocessed content, segmented text chunks, and AI generated training examples such as question–answer pairs, summaries, or labeled data. The system also records processing timestamps, workflow stages, and dataset versions, ensuring traceability and structured record maintenance.

Processing Metadata and Logs

During each stage of data curation, the system collects metadata such as document status, task execution results, quality scores and error logs. These logs help monitor system performance, identify processing issues and ensure transparency in data generation. Metadata plays an important role in debugging, optimization, and future system improvements.

Data Integration

The Training Data Curation Bot integrates multiple internal modules, including document loaders, preprocessing engines, AI task generators, quality evaluation modules and dataset export components. These modules exchange data seamlessly to maintain a continuous and synchronized data flow. The system can also be extended to integrate with external AI platforms, cloud storage services, or machine learning pipelines.

Data Validation and Quality Control

To ensure accuracy and reliability, the system validates uploaded documents and user inputs before processing. Unsupported formats, duplicate files, or invalid configurations are automatically flagged. Quality evaluation mechanisms assess generated training data for relevance, consistency, and usability before inclusion in final datasets, ensuring high data quality standards.

Privacy and Security
All collected data is handled securely in accordance with data protection best practices. Uploaded documents, generated datasets, and metadata are stored securely with controlled access. Authentication and role-based permissions ensure that only authorized users can access or manage sensitive data, maintaining confidentiality and data integrity.

Continuous Data Collection
Data collection is a continuous and dynamic process within the system. As new documents are uploaded and processed, the database is continuously updated with fresh datasets, logs, and quality metrics. This real-time data handling ensures that the system remains up to date and supports efficient retrieval of historical data for analysis, auditing, and reporting. The structured data collection process enables the system to generate insights such as processing trends, dataset usage patterns, and performance metrics, contributing to ongoing system enhancement.

IX. SOFTWARE DESCRIPTION

Python:
Python is the core programming language used to develop the Training Data Curation Bot. It handles backend logic such as document processing, text extraction, preprocessing, AI task execution, and dataset generation. Python ensures efficient, scalable, and reliable processing of large volumes of data.

Web Framework (Flask / Stream lit):
The web framework is used to build the user interface of the system. It provides features for document upload, workflow monitoring, and dataset download through a clean and interactive dashboard. The framework enables seamless communication between the frontend and backend components.

AI / NLP Libraries:
AI and Natural Language Processing libraries are used to extract, process, and transform unstructured text into structured training data. These libraries support tasks such as text cleaning, chunking, and generation of question–answer pairs or summaries.

Database / Storage System:
The database or storage system is used to store uploaded documents, processed text, generated datasets, and metadata. It ensures data integrity, efficient retrieval, and secure management of training data throughout the system lifecycle.

Development Environment:
The development environment provides tools and libraries required to build, test, and run the application locally. It supports debugging, dependency management, and smooth integration of all system components during development.

X. SCREENSHOTS

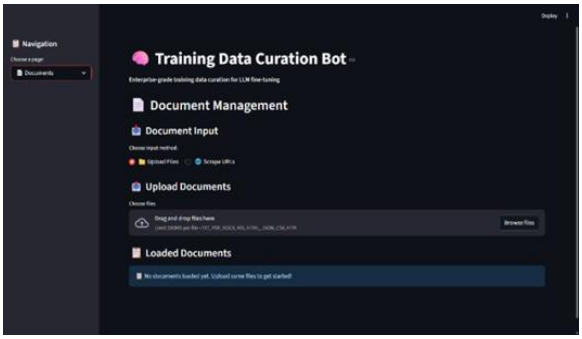


Fig 1.1 Home Page

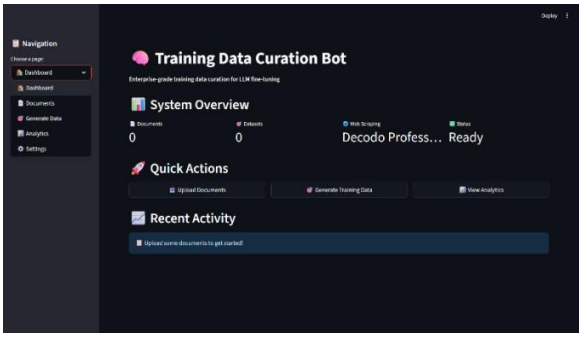


Fig 1.2 Document uploader

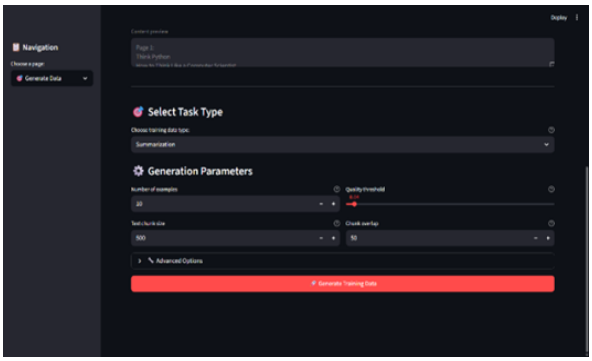


Fig 1.3 Threads hot setter

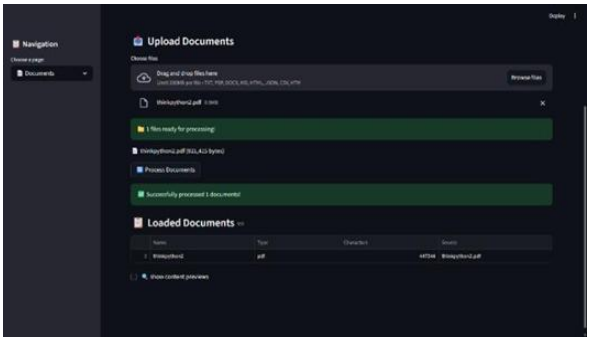


Fig 1.4 Document processing

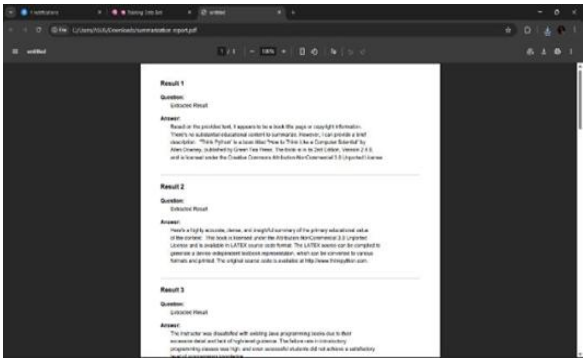


Fig 1.5 Generated pdf

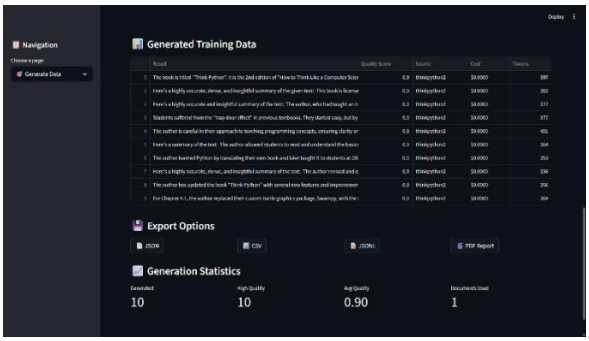


Fig 1.6 Training data

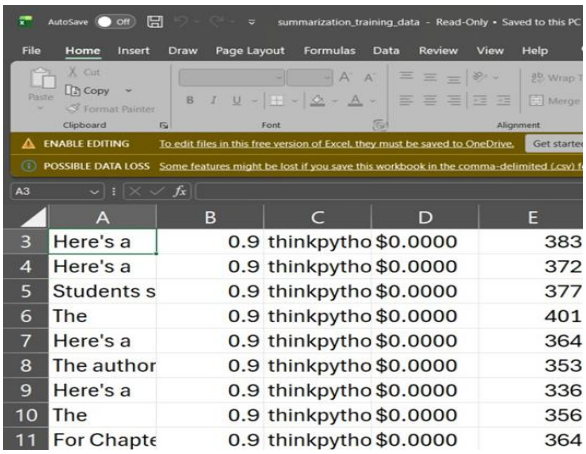


Fig 1.7 Data generated in excel

XI. CONCLUSION

The Training Data Curation Bot represents a significant advancement in automating one of the most critical stages of artificial intelligence development—training data preparation. It bridges the gap between unstructured data sources and AI-ready datasets by replacing manual, repetitive, and error-prone processes with an intelligent, automated pipeline. The system provides a streamlined and centralized platform where users can upload documents, process data, and generate structured training datasets efficiently. Its intuitive web interface allows users to interact easily with the system, while backend automation handles complex tasks such as text extraction, preprocessing, AI-driven data generation, and quality evaluation. The solution enhances efficiency, accuracy, and consistency in dataset creation, enabling faster AI model development and experimentation. Key features such as automated document processing, structured data generation, quality checks, and role-based access control ensure reliability and secure handling of sensitive information. Real time processing updates and transparent workflows help users track progress and maintain confidence in the generated datasets. Developed using Python and modern AI libraries, the system offers a scalable and robust foundation that can adapt to evolving AI requirements. By transforming training data preparation into an automated, intelligent, and scalable workflow, the Training Data Curation Bot significantly reduces manual workload and improves productivity. It supports digital transformation in AI development by promoting standardized data practices, efficient resource utilization, and secure data management. Ultimately, the system contributes to more reliable, transparent, and efficient artificial intelligence solutions, benefiting developers, organizations, and the broader AI ecosystem.

REFERENCES

[1] Alshamrani, A. (2023). Digital transformation of data management systems using web-based platforms. *Journal of Information Technology Management, 34*(2), 45–60.

[2] Breck, E., Polyzotis, N., Roy, S., Whang, S., & Zinkevich, M. (2017). The ML test score: A rubric for ML production readiness. *IEEE Big*

Data Conference.

[3] Chen, X., Liu, Y., & Zhao, W. (2024). Automated data curation for machine learning: Challenges and opportunities. *Journal of Artificial Intelligence Research, 69*, 345–372.

[4] Gartner. (2024). *Trends in AI data preparation and automation platforms.* Gartner Research Report.

[5] Google Cloud. (2023). *Building scalable AI data pipelines with cloud-based workflows.* Technical Documentation.

[6] IBM. (2024). *AI data management and governance for enterprise-scale systems.* IBM Research White Paper.

[7] Kelleher, J. D., & Tierney, B. (2018). *Data preparation for machine learning.* MIT Press.

[8] Khan, M., & Rehman, A. (2023). Document-based dataset generation for NLP applications. *Journal of Big Data, 10*(88).

[9] Knaflic, S., & Dumas, M. (2022). Workflow automation for data-intensive AI applications. *Information Systems, 103*, 101871.

[10] Li, Y., Zhou, T., & Chen, P. (2024). AI-driven data generation and evaluation for scalable model training. *Artificial Intelligence Review, 57*(3), 1–25.

[11] Mhlanga, D. (2024). Digital transformation and automation in data-driven systems. Technology in Society, 78, 102421.

[12] Nguyen, T., Tran, D., & Pham, H. (2025). Smart AI pipelines for document-based training data curation. International Journal of Advanced Computer Science and Applications, 16(1), 210–218.

[13] OpenAI. (2023). Best practices for preparing training data for large language models. Technical Report.

[14] ProcessMaker. (2023). Workflow automation for data processing and AI model readiness. ProcessMaker Blog.

[15] Rahman, M., & Islam, S. (2024). Role-based access control for secure AI data management systems. International Journal of Computer Applications, 185(12), 1–7.

[16] Red Hat. (2024). Automating data pipelines for AI and machine learning workflows. White Paper.

[17] Singh, R., & Patel, V. (2025). Web-based platforms for automated AI dataset creation. International Journal of Emerging Technologies and Innovative Research, 12(2), 56–63.

[18] Sculley, D., et al. (2015). Hidden technical debt in machine learning systems. Advances in Neural Information Processing Systems (NeurIPS).

[19] Zhang, H., Wang, L., & Sun, J. (2023). Intelligent document processing for AI training data generation. IEEE Access, 11, 88214–88227.

[20] Zhao, Q., & Xu, R. (2022). Secure and scalable web-based systems for AI dataset management. Computers & Security, 115, 102604.