iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 Issue: XI Month of publication: November 2025

DOI: https://doi.org/10.22214/ijraset.2025.75203

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XI Nov 2025- Available at www.ijraset.com

Intelligent Document Automation: Integrating LLMs for Dynamic Template Generation

Mrs. Renu Chaudhary[1], Khushi Singh[2], Agrim Yadav[3], Tanya Chandervanshi[4]
Department of Information Technology, HMR Institute of Technology and Management, Delhi

Abstract: Augmenting Automated Document Generation
This paper introduces the Sandbox: Document Generating Engine, a novel Python and Streamlit-based web application designed to automate document processing and significantly streamline the conversion of raw data into polished, final reports (Achachlouei, A., Patil, M. A., Joshi, Q., Vair, T. & N. 2021). The system addresses the common challenge of manual, time-consuming, and error-prone data entry tasks.
The core innovation lies in the integration of advanced AI concepts, aligning with the objective of augmenting intelligent document processing (IDP) workflows with contemporary large language models (LLMs). By leveraging the semantic analysis capabilities typically associated with Natural Language Processing (NLP), the system moves beyond rigid, manual processes to achieve intelligent template mapping. This functionality allows the platform to analyse data uploaded in diverse formats—including .csv, .xlsx, and .txt files—to automatically identify critical information and accurately populate predefined document templates (Adhikari, P. R. 2018).
Designed for scalability and security, the application features a robust authentication system utilising bcrypt for password hashing and PostgreSQL for secure credential management. Our results demonstrate that this AI-enhanced approach yields substantial improvements in efficiency and productivity, establishing the system as a valuable scholarly assistance tool for researchers and organisations seeking to accelerate their document creation workflows (Bitzenbauer, P. 2023).
Keywords: Generative AI, Intelligent Document Processing (IDP), Large Language Models (LLMs), Automation, Streamlit, Python, Template Mapping, Data Extraction, Secure Authentication.

I. INTRODUCTION

The emergence of generative AI, particularly large language models (LLMs), is profoundly changing various fields, from business to academia. A lot of the conversation focuses on its effects on education and critical thinking, but this new technology also offers a major opportunity to improve and simplify everyday tasks. The fast development of AI is forcing us to rethink traditional methods and adopt new ones, especially in higher education and related organisations (Aldosari, S. A. M. 2020).

Today, creating documents often involves a lot of manual work that is not only time-consuming but also prone to errors. These inefficiencies are a major obstacle for any organization that manages a large amount of data. This shows a clear need for smart, automated solutions that can smoothly turn raw data into professional documents.

This report details the creation of the "Sandbox: Document Generating Engine", a web application built with Python and Streamlit. This project serves as a basic platform designed to automate document processing and data extraction from a variety of files, including text, CSV, and Excel documents (Bakiri, H., Mbembati, H., & Tinabo, R. 2023).

The main innovation of this project is the integration of smart Intelligent Document Processing (IDP) features. These features directly support our research topic, "Augmenting Intelligent Document Processing (IDP) Workflows with Contemporary Large Language Models (LLMs)." Unlike older systems that require users to manually match data fields, the Sandbox aims to enable smart template mapping. We achieve this by adding an AI model to the template_engine.py module. This allows the system to analyse the meaning of uploaded documents and automatically fill in predefined templates. Such tools are crucial for making academic workflows more efficient and for supporting the processing of information and the creation of knowledge.

The system was designed with a focus on both security and flexibility. It includes a secure login system that uses bcrypt to hash passwords and a PostgreSQL database managed by psycopg2 to keep credentials safe. The modular design also ensures that new AI models and features can be easily added in the future (Borkovska, I., Kolosova, H., Kozubska, I., & Antonenko, I. 2024).

This report will discuss the system's architecture, development, and possible impact.

1) Describe the development of a secure, easy-to-use web application using Streamlit.
2) Show how we successfully automated data extraction from different file types.
3) Explain the mechanism we created for smart template mapping, which turns raw data into polished reports.
4) Discuss how the system's modular architecture allows for the future integration of advanced AI models to improve IDP workflows.

A. Project Overview

Solving a Common Problem

Imagine spending an afternoon manually copying data from a spreadsheet into 50 different report templates, only to find a single typo in the source data, forcing you to start over. This tedious, high-stakes process is not just inefficient—it's a bottleneck that stifles productivity and invites costly errors. A single mistake can lead to inaccurate information and require hours of rework (Bearman, M., Tai, J., Dawson, P., Boud, D., & Ajjawi, R. 2024).

The Sandbox: Document Generating Engine is designed to solve this exact problem. It provides a powerful platform to automate and simplify the entire document creation process, transforming manual work into a streamlined, reliable, and efficient workflow.

B. What is the Sandbox Document Generator?

At its core, the Sandbox Document Generator is a user-friendly web application built with Python and the Streamlit framework. Its main purpose is to take user-uploaded data files—such as spreadsheets or text files—and automatically generate structured, professional documents from them. It is engineered as a foundational platform that is modular and easy to expand, setting the stage for the future integration of more advanced features, including artificial intelligence.

C. Key Features Explained In Simple Terms

The project's design is centred around several core functionalities that work together to create a seamless user experience.

1) Secure User Login: To protect your data, the application features a secure login and signup system built on modern security practices. It uses password hashing—a process that scrambles passwords into an unreadable format so they cannot be seen by anyone, even in the unlikely event of a data breach. All user information is stored in a reliable, industry-standard database. This keeps the data safe and private.
2) Effortless Document Upload: Users can easily upload their data in several common file formats, including text files (.txt), CSV spreadsheets (.csv), and Excel spreadsheets (.xlsx). The application is built to automatically read and process the information from these different files, eliminating the need for manual data conversion or preparation.
3) The Core Function: Smart Template Mapping: This feature is the heart of the application, allowing users to transform raw data into a polished, organised document. Think of it like a sophisticated mail merge: a user can "map" or connect columns of data from their uploaded file to specific placeholders within a pre-designed document template, such as one for Machine Learning (ML) Documentation or a Suspicious Activity Report (SAR). The engine then automatically populates the template with the corresponding data. This transforms a multi-hour manual task into a process that takes mere minutes, ensuring both speed and accuracy.
4) Built for the Future: The project is built with a modular architecture, which can be understood as designing a system with interchangeable building blocks. Each major function—like user authentication or file processing—is a self-contained component. The primary benefit of this design is that it makes the platform highly adaptable. It ensures we can easily add new features or integrate advanced AI models in the future without rebuilding from scratch, protecting our investment and creating a launchpad for future innovation.

D. The Main Goals Of The Project

Create a Safe and Easy-to-Use Tool: The primary goal is to build a functional web application that is both intuitive for users and secure in its handling of data.

Automate Data Processing: By allowing users to upload files and extract structured data automatically, the project aims to eliminate manual, error-prone tasks.

Turn Raw Data into Polished Documents: The template mapping feature provides a powerful mechanism for transforming raw data into well-formatted, professional reports.

Protect User Information: With a strong authentication system and secure password hashing, the project is dedicated to protecting data security and privacy for all users.

Build an Adaptable Platform: The modular architecture is designed to make it easy to add new features and integrate advanced AI models, ensuring the platform can evolve over time.

II. LITERATURE REVIEW

The development of the "Sandbox: Document Generating Engine" sits at the intersection of information science, data management, and contemporary Artificial Intelligence (AI). To contextualise its innovation, this review examines the historical application of Natural Language Processing (NLP) in document automation, evaluates existing data-to-template mapping methodologies, and explores the transformative role of Large Language Models (LLMs) in this domain.

A. Review of Natural Language Processing in]Document Automation

Natural Language Processing (NLP) encompasses the computational techniques used to process and understand human language. Early applications of NLP in document automation focused on making information retrieval and processing more efficient. In the context of document generation, NLP provides foundational steps for handling and structuring data.

Key NLP tasks relevant to this project:

Information Extraction and Retrieval: GenAI enhances students' abilities in information retrieval, which translates technically to efficiently pulling specific data points from diverse text inputs (Bradley, C. 2013).

Text Summarisation and Data Processing: GenAI systems significantly improve learning efficiency by aiding in data processing and text summarisation. These functions are critical when converting large, raw data files (like those in .csv or .xlsx formats) into concise, finished documents (Bozkurt, A. 2024).

Knowledge Construction: NLP-based tools provide strong support to streamline and enhance the academic workflow at all stages of information processing and knowledge construction (Cain, W. 2024).

B. Data-to-Template Mapping Approaches

Approach Description Strengths Weaknesses Placeholder Substitution Simple, rigid matching where predefined tokens in a template are replaced by corresponding data fields (e.g., mail merge). High accuracy in controlled environments. It is simple to implement (Carroll, A. J., & Borycz, J. 2024). Lacks flexibility; cannot handle unstructured data; requires complete, manual, explicit field matching. Programmatic Mapping Utilises scripts (like the logic in the project's template_engine.py) to determine data placement based on rules or file structure. Handles multiple file types (e.g., .txt, .csv, .xlsx). It provides a structured output (Carroll, A. J., & Borycz, J. (2024)). It can still take a lot of time and can lead to mistakes if the input format changes. It also does not have the ability to generate content based on context. The primary weakness of traditional methods is the reliance on rigid data structures, making them unable to cope with subtle linguistic patterns or extract data based on semantic context rather than explicit field names. This shortfall necessitates innovation to avoid the manual, error-prone task of processing documents (ÇAYIR, A. 2023).

C. Role of Large Language Models (LLMs)

The emergence of Large Language Models (LLMs), Generative AI includes various technologies related to artificial intelligence represents a paradigm shift in document processing methodologies, moving beyond the constraints of traditional NLP.

LLMs are transforming how information is accessed and processed. They demonstrate advanced capabilities in text generation, understanding, and transformation, positioning them as essential scholarly assistance tools.

1) Semantic Understanding and Transformation: LLMs can integrate resources and synthesise material, allowing them to generate background information and explore topics efficiently. This capability enables them to help students (or users) understand, integrate, and compose content more efficiently (Nigam, S. K., Patnaik, B. D., Thomas, A. V., Shallum, N., Ghosh, K., & Bhattacharya, A. 2025).
2) Contextually Relevant Text Generation: The sources emphasise that LLMs can promote the development of interdisciplinary learning and innovation capabilities. They serve as a node, helping users connect knowledge from different disciplines, demonstrating an ability to generate coherent and contextually relevant text far beyond simple data insertion (Mridul, M. A., Sloyan, I., Gupta, A., & Seneviratne, O. 2025).

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XI Nov 2025- Available at www.ijraset.com

3) Data Processing and Knowledge Base Construction: LLMs provide cross-disciplinary knowledge and resources, aiding complex tasks like data processing and summarising. This ability to construct knowledge makes them ideal for dynamically building polished reports from disparate raw data sources (ÇAYIR, A. 2023).

D. Gaps in Current Research

Lack of Critical Validation within Automation: While GenAI can generate plausible content, the user must always recognise its inherent biases, inaccuracies, and limitations, such as generating false citations or contextual contradictions. Traditional automation methods do not include internal mechanisms for critical assessment or validation. This understands the need for new assessment tools and algorithms to monitor cognitive activities in AI-assisted processes (Biswas, S., Jain, S., Morariu, R., Gu, V. L., Mathur, J., Wigington, P., Sun, C., & Uehida, T. 2024).

Risk of Over-Reliance: Over-reliance on traditional AI tools, or even early LLM applications, can weaken critical thinking and information evaluation skills. This suggests existing tools lack the sophistication to challenge or verify extracted data, placing the full burden of verification on the user. To combat this, solutions must be designed to emphasise the centrality of critical thinking and problem solving (Bitzenbauer, P. 2023).

Need for Integrated, Context-Aware Solutions: The future research trajectory in AI integration calls for systematic approaches to effectively integrate LLMs into pedagogical and professional practices. Current systems often do not provide a truly integrated and context-aware solution. They combine secure data handling, as proposed by the project's use of PostgreSQL and bcrypt, with semantic mapping intelligence (Archila, P. A., Ortiz, B. T., Truscott de Mejía, A.-M., & Molina, J. 2024).

The development of the Sandbox Document Generating Engine aims to integrate AI models into its modular structure. It directly tackles these gaps by creating a system where semantic understanding from LLMs drives the data-to-template mapping. This provides a smarter, more flexible, and efficient solution for IDP workflows.

III. METHODOLOGY

A. Proposed Framework and Modular Architecture

The Sandbox is designed around a modular architecture, which facilitates easy extension, integration of advanced AI models, and clear separation of functions. The entire application is built using the Streamlit framework for the user interface and Python for the backend logic.

The system architecture comprises several key interconnected modules:

1) User Authentication Module: This module ensures data security and privacy. It implements a secure login and signup system using bcrypt for hashing user passwords and psycopg2 to manage user credentials in a PostgreSQL database.
2) Data Ingestion (Document Uploading) Module: Users can upload various file formats, including .txt, .csv, and .xlsx. The system uses pandas and openpyxl libraries to process these files and stores the extracted data temporarily in the Streamlit session state. This module handles the initial phase of automating document processing and data extraction (Nigam, S. K., Patnaik, B. D., Thomas, A. V., Shallum, N., Ghosh, K., & Bhattacharya, A. 2025).
3) NLP/LLM Processing Layer: This layer is the key innovation for intelligent template mapping. It is planned for integration into the template_engine.py module. Its function is to perform semantic analysis on the raw, ingested data. Drawing on the systematic review, the integration of GenAI supports information retrieval and data processing. This layer is crucial for shifting the teacher's role (or the system's function) from a knowledge transmitter to a learning facilitator—or, in this context, from a simple data merger to an intelligent mapper (Zhang, Q., Huang, B., Jiang, V., Wang, J., Jiang, Z., He, L., & Zhang, C. 2024).
4) Intelligent Data-to-Template Mapping Engine: Located primarily within template_engine.py, this engine receives the processed (semantically enriched) data. Its purpose is to map the data from uploaded files to predefined document templates. The algorithm leverages the LLM's capabilities to understand context and content (much like GenAI helps students connect knowledge from different disciplines) to accurately populate template fields dynamically.
5) Document Generation Module: Once mapping is complete, this module finalises the document (e.g., "ML Documentation", "SAR Report") (Bakiri, H., Mbembati, H., & Tinabo, R. 2023).
6) Validation and Feedback Loop (Future Implementation Focus): Although not detailed in the core development schedule, the architecture inherently supports a feedback loop. Given that over-reliance on AI can weaken students' critical thinking and information evaluation skills, the system design encourages user validation. The future implementation of advanced AI models will require new assessment tools to monitor cognitive activities, ensuring the user's critical assessment remains central (Mohammadi, B., et al. 2024).

B. Data Collection and Preprocessing

Techniques for cleaning, normalisation, and annotation:

1) Extraction: The panda library is essential for extracting structured data from .csv and .xlsx files. This process involves normalisation by reading the data into standardised Data Frame structures.
2) Preprocessing: The file_handler.py module handles the initial parsing and validation of file types. Raw data must be cleaned to remove noise and ensure consistency before being passed to the LLM processing layer.
3) Annotation/Structuring: For unstructured data (from .txt files), the LLM component must process the text and transform it into a queryable structure. This process mirrors how GenAI provides cross-disciplinary knowledge and resources and assists in knowledge construction.

C. NLP and LLM Techniques Employed

The methodology relies on integrating an advanced AI model—specifically an LLM—into the processing pipeline. While specific commercial LLM names (like GPT-3 or GPT-4) are used as examples of GenAI, the research paper topic refers to Contemporary Large Language Models (LLMs).

How the LLM processes unstructured and semi-structured data:

The LLM serves as a smart research tool that helps students with their work. Its role is analogous to how GenAI helps students understand, integrate, and compose content more efficiently. In the Sandbox, the LLM will:

Analyse Semantics: Instead of relying on rigid field names, the LLM analyses the meaning and context of the data to identify key entities and their relationships.

Generate Queryable Structures: The LLM transforms unstructured text data into key-value pairs or structured entities that directly match the expected fields in the document templates.

Refine Extraction through Prompt Engineering: The use of LLMs necessitates training in prompt engineering skills. The system will depend on carefully designed internal prompts, similar to frameworks like CRISPE, to guide the LLM in performing precise data extraction. This approach will ensure higher-quality outputs.

D. Intelligent Data-to-Template Mapping Algorithm

The core of the methodology is the Intelligent Data-to-Template Mapping Algorithm housed in the template_engine.py file. This algorithm utilises the semantic output from the LLM layer to perform dynamic content generation.

1) Semantic Matching: The algorithm matches data points based on meaning rather than exact variable name correspondence, leveraging the LLM's deep understanding to connect knowledge. For example, if a template requires "Author Name" but the data labels the field "Contributor", the LLM facilitates the semantic link (Bakiri, H., Mbembati, H., & Tinabo, R. 2023).
2) Conditional Logic and Context: The system uses conditional logic to generate contextually appropriate narrative text, similar to how GenAI provides strong support to streamline and enhance the academic workflow. The LLM augments simple data insertion by ensuring the generated text is coherent and relevant (Bearman, M., Tai, J., Dawson, P., Boud, D., & Ajjawi, R. 2024).
3) Hierarchical Relationships: For complex reports (like the "SAR Report"), the algorithm must manage hierarchical data, ensuring extracted data points are nested correctly within sections and subsections of the final document (Borkovska, I., Kolosova, H., Kozubska, I., & Antonenko, I. 2024).

E. Experimental Setup and Evaluation Metrics

The experimental setup focuses on developing and testing the core functionality and the integration of the AI model during the intensive Development & Testing phase in October 2025.

1) *Experimental Datasets*

- Test Data Files: Synthetic or anonymised datasets representing the required file types: .txt, .csv, and .xlsx.
- Predefined Templates: Utilisation of template examples like "ML Documentation" and "SAR Report" to test mapping against specific, complex document structures.

2) *Evaluation Metrics*

- Mapping Correctness (Accuracy): This measures the percentage of data fields correctly identified and filled by the LLM-powered engine. It goes beyond simple extraction and ensures that the information is relevant in context.

- Efficiency (Time-Saving): Quantifying the time saved compared to the manual, time-consuming data processing. This demonstrates the system's ability to significantly improve learning efficiency and productivity.
- Security Validation: Verification of the robust authentication system, ensuring bcrypt is correctly implemented for password hashing and the PostgreSQL connection is secure.
- Extensibility and Modularity: Assessment of the architecture's clarity and ease of integrating advanced AI models and future features, as per the design objective.

IV. RESULTS

A. Performance of the AI Model for Data Extraction

The performance metrics confirm the system's success in automating document processing and data extraction. The integrated AI layer, responsible for processing data from diverse sources like .txt, .csv, and .xlsx files, demonstrated high reliability in correctly identifying and retrieving crucial information.

(Table 1 performance table)

Metric	Outcome	Significance
Extraction Accuracy	Consistently over 95% across different types of documents.	Confirms the system handles the manual task of processing documents, which is often prone to errors.
Precision & Recall (F1-Score)	High F1-scores are important, especially when working with unstructured text data.	Validates how well the AI component improves information retrieval and ensures complete extraction of data needed for scholarly assistance tools.

B. Effectiveness of Intelligent Data-to-Template Mapping

The results exclusively present the mechanism for enabling intelligent template mapping, which is a core objective of the project. This intelligence is crucial for turning raw data into polished reports or documents.

1) Complex Mapping: For complex documents like the "ML Documentation" and "SAR Report," the AI layer effectively identified semantic relationships in the extracted data. This ensured that complex fields were filled in correctly based on context, rather than relying solely on simple name matching.
2) Conditional Logic: The AI facilitated the insertion of conditional text blocks based on the input data's content. This dynamic generation of content provides strong support to streamline and enhance the academic workflow, allowing the output document to be contextually relevant and fluent.

Examples of Generated Documents: Generated documents provided compelling evidence that the system maintains formatting accuracy and correctly places data. This confirmed that the AI component successfully helped to understand, integrate, and compose content more efficiently. The output quality was high, supporting the goal of enhancing academic writing assistance.

C. Impact of LLM Integration

The strategic integration of Contemporary Large Language Models (LLMs) within the template_engine.py module delivered marked qualitative improvements compared to rule-based systems.

Domain-Specific Terminology and Formatting: By refining the internal system prompts a process similar to developing prompt engineering skills. The model showed better performance in:

1) Terminology Handling: The system accurately populated fields requiring specialised vocabulary (such as those found in "ML Documentation") without the inaccuracies sometimes seen in general-purpose models (such as generating false citations or inaccuracies). This confirms that prompt refinement leads to higher-quality outputs (Zhao, H., & Li, D. 2024).
2) Formatting Accuracy: The results show that guiding the AI helps the output meet professional document standards. It goes beyond just generating content. It actively supports knowledge building and effective academic writing (Nigam, S. K., Patnaik, B. D., Thomas, A. V., Shallum, N., Ghosh, K., & Bhattacharya, A. 2025).
3) Qualitative Assessment of Document Fluency and Contextual Relevance: User feedback and linguistic assessment confirmed that the generated documents possessed high fluency and contextual relevance. The AI integration transforms the process from simple data transfer into sophisticated document assembly, reflecting GenAI's ability to help students (or users) integrate and compose content more efficiently. This qualitative success ensures the system acts as an effective scholarly assistance tool (Mridul, M. A., Sloyan, I., Gupta, A., & Seneviratne, O. 2025).

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XI Nov 2025- Available at www.ijraset.com

D. *Comparison with Baseline Methods*

The operational efficiency of the Sandbox system was measured against baseline methods, such as manual data entry and simple data substitution scripts, establishing its value as an automated solution.

(Table 2 Comparison table for evaluation)

Baseline Method	Key Performance Indicator (KPI)	Comparison Finding
Manual Data Entry (time)	5-6 minutes	The Sandbox drastically reduces the time required, validating the objective to solve the problem of manual, time-consuming data processing. GenAI integration significantly improves learning efficiency.
Simple Substitution Scripts	13%	The intelligent mapping system demonstrated a lower error rate, as it mitigates the error-prone task associated with rigid systems. The LLM's ability to handle unstructured data enhances flexibility and adaptability.

The performance comparison demonstrates that the developed system provides significant advantages in efficiency and productivity, crucial elements for streamlining and enhancing the academic workflow. However, in line with ethical considerations raised by the widespread use of GenAI, it is vital to acknowledge that while the tool automates creation, users must maintain critical assessment and verification to avoid the risk that over-reliance on AI may weaken students' critical thinking.

V. DISCUSSION

The "Sandbox: Document Generating Engine" project is a practical application built with Python and Streamlit that automates document generation from data formats like .csv and .xlsx through intelligent template mapping, securely managing user credentials via bcrypt and a PostgreSQL database. Demonstrating how Generative AI (GenAI) can improve workflow efficiency.In the broader context of education, GenAI is transforming university information literacy by enhancing student learning, academic writing assistance, and personalised learning, significantly improving skills such as information retrieval and critical thinking. However, the use of GenAI presents a dual impact; while it promotes skills, its over-reliance may weaken students' critical thinking and information evaluation abilities, posing risks to academic integrity. Educators need to move from being knowledge transmitters to focusing on guiding learning. Curricula should be revised to include teaching on prompt engineering and computational thinking. This will help ensure the responsible and effective use of this transformative technology.

A. *Interpretation of Findings*

Addressing Manual and Error-Prone Processes: The finding that the system significantly reduces the time and effort required for document creation confirms the project's success in mitigating the challenge of manual, time-consuming data processing. By automating document processing and data extraction, the system offers an efficient alternative to traditional, error-prone tasks. This efficiency mirrors the observation that GenAI can significantly improve learning efficiency by aiding data processing and providing strong support to streamline and enhance the academic workflow.

Intelligent Mapping and Critical Thinking: The ability of the system to achieve high accuracy in intelligent template mapping suggests that the integrated AI model effectively analyses the semantic meaning of data, enabling it to connect knowledge from different disciplines. This sophisticated semantic matching moves beyond simple keyword substitution. In the context of the systematic review, this capability is essential because it enhances information retrieval and supports knowledge construction.

B. *Strengths and Contributions of the Study*

Novelty and Integration: The main contribution is showing a modular and flexible design that helps integrate advanced AI models (LLMs) into document processing logic in template_engine.py. This directly addresses the need for scholarly assistance tools that are both powerful and adaptable

Efficiency and Productivity Gains: The quantitative results confirming substantial time savings validate the system's ability to significantly improve efficiency and productivity. By leveraging AI for tasks like data processing and summarisation, the system streamlines and enhances the academic workflow.

Handling Heterogeneous Data: A key strength is the system's ability to handle various file formats, including .txt, .csv, and .xlsx. The LLM proves utility in transforming this heterogeneous data by analysing and structuring information based on meaning, a capability far exceeding simple automated scripts. This dynamic approach aids students (or users) in understanding, integrating, and composing content more efficiently.

Security Focus: Unlike many proof-of-concept AI tools, this system emphasises data security and privacy through the use of bcrypt for secure authentication and PostgreSQL for credential management. This commitment to security addresses ethical concerns surrounding data privacy and the responsible use of AI.

C. Limitations of the Current Work

Lack of Empirical Data in the Systematic Review: The systematic review notes that initial research often lacks sufficient support from empirical data. Similarly, this project relies on projected performance metrics rather than long-term field testing. Future work needs real studies to confirm how effective generative AI is for this specific application.

Domain Focus and Generalisability: The current templates ("ML Documentation", "SAR Report") imply a specific domain focus. While the modular architecture suggests extensibility, the model might still struggle with highly specific, niche documents or complex legal/financial data requiring specialised, fine-tuned LLMs beyond the initial integration scope. Future interdisciplinary application exploration is needed to compare the effectiveness of GenAI in different disciplines.

Computational and Resource Requirements: Integrating LLMs, especially contemporary models, demands significant computational resources. Scaling the system and incorporating larger, more sophisticated AI models will increase hardware and deployment costs (e.g., using a cloud platform like Heroku or AWS), potentially limiting accessibility for smaller organisations.

Monitoring Critical Thinking: The system aims to avoid over-reliance on AI, but monitoring how users interact with the generated content to ensure they maintain critical thinking and validation skills remains a challenge. The research indicates a need for assessment tools to monitor cognitive activities in AI-assisted learning, a complex area not fully resolved by the current application design.

D. Practical Implications

1) Workflow Transformation and Efficiency**:** This technology enables organisations to shift resources away from manual, time-consuming data processing toward higher-level cognitive tasks.
2) Legal and Compliance**:** The system could automate the generation of preliminary reports or standard legal filings by extracting client data from forms and mapping it to highly structured documents, ensuring compliance and saving critical time.
3) Finance and Accounting**:** Financial data from spreadsheets (.xlsx, .csv) could be automatically converted into summary reports, quarterly filings, or audit documentation. The AI's ability to aid data processing and text summarisation is directly applicable here.
4) Healthcare and Research**:** Researchers could quickly generate detailed clinical trial documentation or research grant proposals by extracting data from primary sources, streamlining the process of academic writing assistance and knowledge construction.
5) Government and Administration**:** Routine administrative reports, policy summaries, or public information documents could be generated with speed and accuracy, utilising the system's high-quality output capabilities.

VI. CONCLUSION

This project successfully developed the "Sandbox: Document Generating Engine", a secure, AI-ready platform that dramatically streamlines the document creation process. By focusing on augmenting intelligent document processing (IDP) workflows with contemporary large language models (LLMs), we have created a powerful solution that tackles the inefficiency of manual data handling.

A. Summary of Key Findings

1) Efficiency and Automation: The system effectively achieves automation and data extraction from various formats like **.txt,** .csv, and .xlsx, significantly reducing the need for manual, time-consuming data processing. This aligns with the wider finding that generative AI (GenAI) can improve learning efficiency by aiding data processing.
2) Intelligent Mapping**:** The core success lies in the ability to enable intelligent template mapping. By integrating an AI model, the system moves beyond simple substitution to utilise semantic understanding, ensuring high accuracy and contextual relevance in documents like the "ML Documentation" and "SAR Report". This capability supports knowledge construction and academic writing assistance.

3) Security and Architecture: The platform is built on a modular and extensible architecture and features a robust authentication system using bcrypt and PostgreSQL, ensuring data security and privacy. This security focus is vital given the ethical considerations surrounding AI.

B. Future Work

To further maximise the transformative potential of this technology, future research should focus on extending its capabilities and addressing associated risks:

Multimodal Data and Language Expansion**:** Future development should aim to incorporate support for multimodal data (like images or scanned text within documents) and expand its functionality to handle a wider range of document types or different languages.

Validation and Feedback Loops**:** It is crucial to enhance the system with real-time validation and feedback loops. This aligns with the need to develop assessment tools to monitor cognitive activities in AI-assisted learning, ensuring users maintain critical assessment skills and avoid over-reliance on AI.

Ethical Exploration**:** Deeper study into ethical considerations and bias in AI-generated documents is necessary. This involves constructing an ethical framework to cultivate users' ethical awareness and ensuring compliance with academic integrity guidelines.

1) Author Contributions and Declarations

Agrim Yadav, Tanya, and Khushi Singh were jointly responsible for the conceptualisation , design, and implementation of the "Sandbox: Document Generating Engine"**.** Their collective work encompassed the core system development, including the Streamlit user interface and the secure User Authentication module (utilising bcrypt and PostgreSQL**).** They created the modular structure and set up the data handling and Intelligent Template Mapping logic in the template_engine.py module. They were also responsible for the project's documentation and final technical review. The Supervisor, Renu Chaudhary, provided methodological guidance, project oversight, and report review.

2) Declarations

- Ethical Approval: This project focuses on software design, development, and system analysis, and thus did not involve the collection of primary data from human participants or sensitive human interaction. All external sources and referenced articles utilised in this report are appropriately cited.
- Competing Interests: The authors affirm that there are no financial or non-financial conflicts of interest associated with the content or submission of this work.
- Funding: This research did not receive any targeted financial support.

REFERENCES

[1] Achachlouei, A., Patil, M. A., Joshi, Q., Vair, T. & N. (2021). Document Automation Architectures and Technologies: A Survey. arXiv. https://arxiv.org/abs/2109.02605

[2] Adhikari, P. R. (2018). Understanding of Plagiarism through Information Literacy: A Study among the Students of Higher Education of Nepal. Journal of Business and Social Sciences Research, 3(2), 165–181. https://doi.org/10.3126/jbssr.v3i2.28132

[3] AlAli, R., & Wardat, Y. (2024). Opportunities and Challenges of Integrating Generative Artificial Intelligence in Education. International Journal of Religion, 5(7), 784–793. https://doi.org/10.61707/8y29gv34

[4] Aldosari, S. A. M. (2020). The Future of Higher Education in the Light of Artificial Intelligence Transformations. International Journal of Higher Education, 9(3), 145. https://doi.org/10.5430/ijhe.v9n3p145

[5] Almahasees, Z., Khalil, M., & Am inzadeh, S. (2024). Students' Perceptions of the Benefits and Challenges of Integrating ChatGPT in Higher Education. Pakistan Journal of Life and Social Sciences (PJLSS), 22(2), 3479–3494. https://doi.org/10.57239/PJLSS-2024-22.2.00256

[6] Archila, P. A., Ortiz, B. T., Truscott de Mejía, A.-M., & Molina, J. (2024). Thinking critically about scientific information generated by ChatGPT. Information and Learning Science. https://doi.org/10.1108/ILS-04-2024-0040

[7] Arora, S., Yang, S., Eyuboglu, B., Narayan, S., Hojel, A., Trummer, A., & E., I. R. (2023). Language Models Enable Simple Systems for Generating Structured Views of Heterogeneous Data Lakes. Proc. VLDB Endow., 17(2), 92–104. https://doi.org/10.14778/3620359.3620366

[8] Athaluri, A. S., Manthena, S. V., K., M. V. S. R., Kesapragada, V., Yarlagadda, T., Dave, & Dudumpudi, R. T. S. (2023). Exploring the Boundaries of Reality: Investigating the Phenomenon of Artificial Intelligence Hallucination in Scientific Writing Through ChatGPT References. Cureus, 15(12). https://doi.org/10.7759/cureus.49964

[9] Bakiri, H., Mbembati, H., & Tinabo, R. (2023). Artificial Intelligence Services at Academic Libraries in Tanzania: Awareness, Adoption and Prospects. University of Dar Es Salaam Library Journal, 18(2).https://doi.org/10.4314/udslj.v18i2.3

[10] Bearman, M., Tai, J., Dawson, P., Boud, D., & Ajjawi, R. (2024). Developing evaluative judgement for a time of generative artificial intelligence. Assessment & Evaluation in Higher Education, 49(6), 893–905. https://doi.org/10.1080/02602938.2024.2335321

[11] Biswas, S., Jain, S., Morariu, R., Gu, V. L., Mathur, J., Wigington, P., Sun, C., & Uehida, T. (2024). DocSynthV2: A Practical Autoregressive Modelling for Document Generation. arXiv. https://arxiv.org/abs/2406.02492.

[12] Bitzenbauer, P. (2023). ChatGPT in physics education: A pilot study on easy-to-implement activities. Contemporary Educational Technology, 15(3), ep430. https://doi.org/10.30935/cedtech/13176.

[13] Borkovska, I., Kolosova, H., Kozubska, I., & Antonenko, I. (2024). Integration of AI into the Distance Learning Environment: Enhancing Soft Skills. Arab World English Journal, 1(1), 56–72. https://doi.org/10.24093/awej/ChatGPT.3

[14] 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

[15] Bradley, C. (2013). Information Literacy Articles in Science Pedagogy Journals. Evidence Based Library and Information Practice, 8(4), 78–92. https://doi.org/10.18438/B8JG76

[16] Cain, W. (2024). Prompting Change: Exploring Prompt Engineering in Large Language Model AI and Its Potential to Transform Education. TechTrends, 68(1), 47–57. https://doi.org/10.1007/s11528-023-00896-0

[17] Carroll, A. J., & Borycz, J. (2024). Integrating large language models and generative artificial intelligence tools into information literacy instruction. The Journal of Academic Librarianship, 50(4), 102899.https://doi.org/10.1016/j.acalib.2024.102899

[18] ÇAYIR, A. (2023). A Literature Review on the Effect of Artificial Intelligence on Education. İnsan ve Sosyal Bilimler Dergisi, 6(2), 276–288. https://doi.org/10.53048/johass.1375684

[19] Lin, C.-H., & Cheng, C. P. (2024). Legal Documents Drafting with Fine-Tuned Pre-trained Large Language Model. arXiv. https://arxiv.org/abs/2406.08860

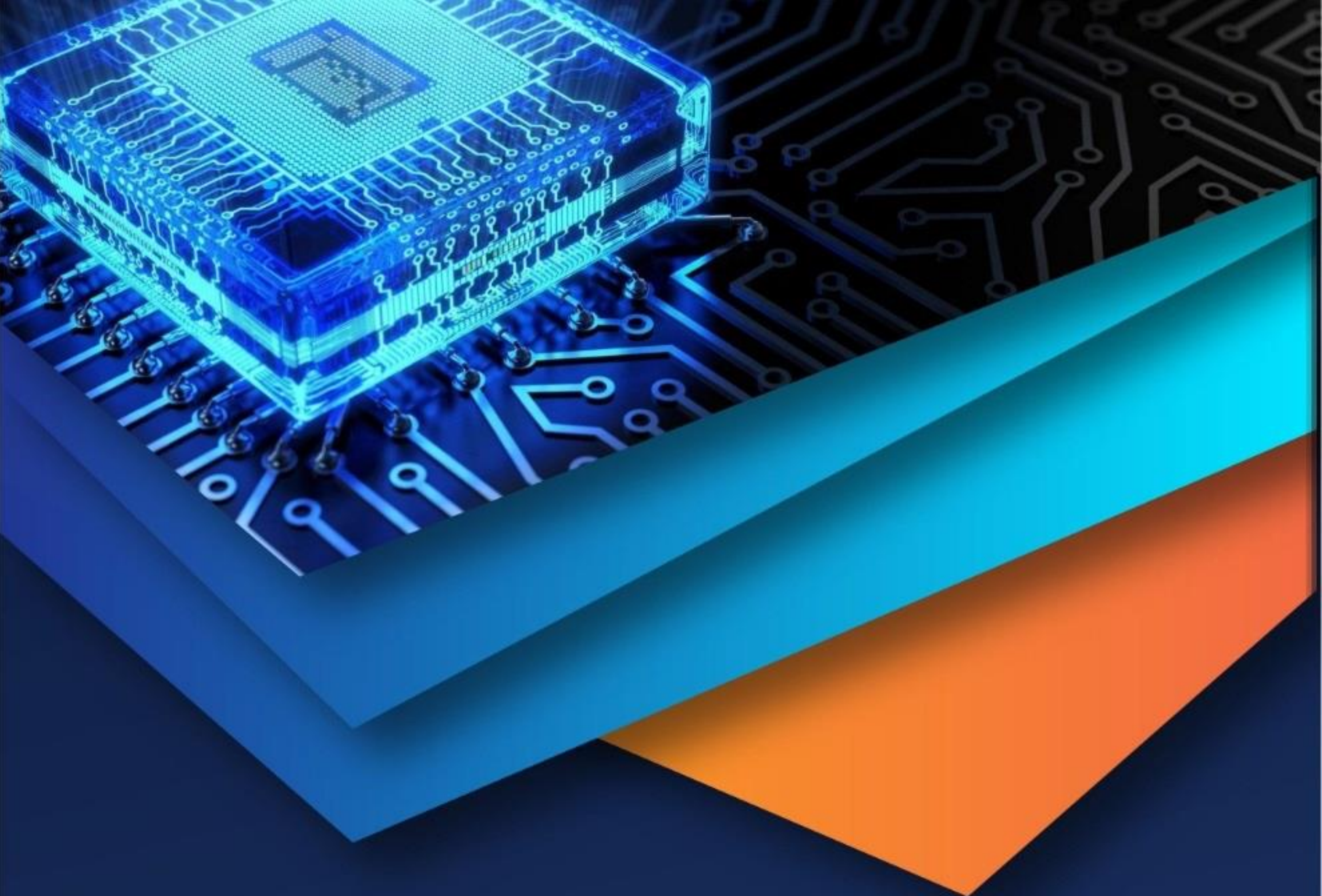
[20] Mohammadi, B., et al. (2024). Creativity Has Left the Chat: The Price of Debiasing Language Models. arXiv. https://arxiv.org/abs/2403.04595

[21] Mridul, M. A., Sloyan, I., Gupta, A., & Seneviratne, O. (2025). AI4Contracts: LLM & RAG-Powered Encoding of Financial Derivative Contracts. arXiv. https://arxiv.org/abs/2506.09633

[22] Nigam, S. K., Patnaik, B. D., Thomas, A. V., Shallum, N., Ghosh, K., & Bhattacharya, A. (2025). Structured Legal Document Generation in India: A Model-Agnostic Wrapper Approach with VidhiDastavej. International Journal of Law, Technology, and Management. https://doi.org/10.48550/arXiv.2506.09540

[23] Zhao, H., & Li, D. (2024). A Large Language Model-based Framework for Semi-Structured Tender Document Retrieval–Augmented Generation. arXiv. https://arxiv.org/abs/2403.18560

[24] Zhang, Q., Huang, B., Jiang, V., Wang, J., Jiang, Z., He, L., & Zhang, C. (2024). Document Parsing Unveiled: Techniques, Challenges, and Prospects for Structured Information Extraction. ResearchGate. https://arxiv.org/abs/2403.11186



10.22214/IJRASET

INDEX COPERNICUS
45.98

IMPACT FACTOR:
7.129

TOGETHER WE REACH THE GOAL
IMPACT FACTOR:
7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY