*Grenze International Journal of Engineering and Technology, June Issue*

# Automating Template-based PDF Customization using Large Language Models

Rohita Yamaganti[1], Naga Siva Jyothi Kompalli[2], R Kameshwar Reddy[3], V. Tharun[4], S. Sunayana[5] and V. Rakshitha[6]
[1-2]Associate professor, Dept of IT Sreenidhi Institute of Science and Technology
Email: rohita.y@sreenidhi.edu.in, sivajyothi.p@sreenidhi.edu.in
[3]Asst. Professor & Dept. of IT Sreenidhi Institute of Science and Technology
Email: kameshwar.r@sreenidhi.edu.in
[4-6]Dept. of IT, Sreenidhi institute of science and technology
Email: tarunvemula26@gmail.com, sadulasunayana@gmail.com, rakshithavydya@gmail.com

***Abstract*— Automating template-based PDF customization is essential across industries such as finance, healthcare, and legal documentation. Traditional scripting-based methods often lack flexibility and scalability. This research investigates the integration of Large Language Models (LLMs) to enhance the generation and modification of structured PDFs. By leveraging techniques such as prompt engineering, Named Entity Recognition (NER), and layout-aware processing, LLMs enable dynamic content mapping onto predefined templates while maintaining coherence and accuracy. Comparative analysis against rule-based approaches highlights LLMs' advantages in adaptability, processing efficiency, and reduced manual intervention. Experimental evaluation demonstrates improvements in content accuracy, formatting precision, and processing latency, validating the potential of LLMs for scalable and intelligent document automation.**

***Keywords*— Large Language Models, PDF Customization, Template- Based Document Generation, Document Automation, Natural Language Processing.**

## I. INTRODUCTION

Automated document generation is a fundamental requirement in enterprise operations, enabling businesses to efficiently handle large-scale document processing. Traditionally, scripting-based methods, such as JavaScript and Python libraries like ReportLab and PyPDF2, have been employed for PDF customization. However, these methods often require extensive rule-based configurations and manual intervention, making them inflexible for dynamic content generation. The growing capabilities of Large Language Models (LLMs) present an opportunity to enhance document automation by leveraging their natural language understanding and contextual reasoning [1]. LLMs offer the ability to dynamically interpret user inputs, extract relevant data, and format structured outputs that can seamlessly integrate with predefined PDF templates. Techniques such as prompt engineering, Named Entity Recognition (NER), and layout-aware processing further enhance the precision and adaptability of LLM-driven document generation systems. Recent advancements in AI-based document automation have demonstrated improved accuracy and efficiency compared to traditional rule-based systems [2,3].

*Grenze ID: 01.GIJET.11.2.798_22*

This paper explores a novel framework utilizing LLMs for customizing template- based PDFs, with a focus on adaptability, reduced processing time, and enhanced personalization. By evaluating performance metrics such as content accuracy, formatting precision, and latency, we aim to demonstrate the efficacy of LLMs in automating document workflows.
Additionally, a comparative analysis against existing template-matching methods underscores the advantages of AI-driven approaches in achieving greater flexibility and efficiency [4].

## II. Literature Survey

Brown et al. (2020) explored the capabilities of GPT-3 in text-based automation, demonstrating its potential in structured content generation. Their study highlights LLMs' ability to generate human-like responses while maintaining contextual coherence, which is crucial in document automation.
The researchers further examined the impact of fine-tuning and prompt engineering techniques in improving accuracy for specialized domains [5].
Rajpurkar et al. (2018) examined Named Entity Recognition (NER) models in extracting structured data from unstructured text, showing that deep learning-based NER significantly improves information retrieval accuracy in document processing applications. Their work emphasized the role of entity linking and context-aware embeddings in enhancing extraction efficiency from legal and financial documents [6].
Xu et al. (2021) investigated prompt engineering techniques to optimize LLM responses for specific domains. Their findings suggest that well-designed prompts enhance output relevance and consistency, which is vital for template-based PDF customization. They also demonstrated how domain adaptation through reinforcement learning can further improve text generation in structured documents [7].
Yin et al. (2022) compared LLM-based document automation with rule-based template-matching approaches, revealing that LLMs provide greater adaptability and reduced development overhead in enterprise applications. Their evaluation metrics included content fidelity, semantic coherence, and system adaptability in handling diverse document structures [8].
Lewis et al. (2020) introduced BART, a transformer-based model that excels in text summarization and structured data generation, which are key components in intelligent document processing. Their research validated BART's ability to reconstruct missing sections in incomplete documents while maintaining syntactic correctness [9].
Devlin et al. (2019) presented BERT, demonstrating its effectiveness in contextual text understanding, which plays a significant role in ensuring accurate data mapping in automated PDF generation. Their study compared BERT's embeddings against traditional NLP models, concluding that bidirectional transformers significantly improve text classification and entity recognition tasks [10].
Yang et al. (2022) proposed an AI-driven legal document processing system utilizing LLMs, showing improvements in contract automation and compliance tracking through intelligent text analysis. Their framework integrated knowledge graphs with LLMs to enhance interpretability and reasoning within complex legal documents [11].
Li et al. (2021) explored layout-aware neural networks for document processing, highlighting their role in accurately aligning text and tabular data in complex templates. They introduced a hybrid architecture combining CNNs with transformers to better understand spatial structures in PDF documents [12].
Shen et al. (2020) examined transformer-based models in financial report automation, demonstrating enhanced accuracy and efficiency compared to traditional scripting approaches. Their research included comparative performance evaluations on financial datasets, showing a notable reduction in processing errors and time complexity [13].
Kim et al. (2021) analyzed multimodal AI systems combining text and visual cues for improved document layout processing, an approach that aligns well with LLM- powered PDF customization. Their study demonstrated that integrating image embeddings with text-based transformers enhances document structure recognition, enabling more accurate template-based formatting [14].
He et al. (2023) developed an AI-based medical report generation framework leveraging LLMs, proving their efficacy in domain-specific structured text generation. Their system incorporated clinical language models fine-tuned on electronic health records, improving medical document synthesis and patient report customization [15].
Zhang et al. (2023) introduced a hybrid LLM approach combining deep learning with knowledge graphs for enhanced document understanding, which can be beneficial in personalized PDF automation. Their research validated the approach through extensive experiments, showing improved contextual comprehension and entity disambiguation in complex text environments [16].

## III. Existing Techniques for Template-Based PDF Customization

Several traditional and modern approaches exist for generating and customizing template-based PDFs. These techniques vary in their complexity, flexibility, and scalability. Below, we discuss key existing techniques and provide justifications using tabular comparisons.

### *A. Rule-Based and Scripting Methods*

Rule-based and scripting approaches rely on predefined templates and manual configurations to insert data into structured PDFs. Libraries such as ReportLab, PyPDF2, and PDFKit allow developers to automate document generation. These methods are widely used in financial and legal documentation, where predefined rules govern document structure. However, they are highly rigid, requiring extensive programming effort for modifications [17].

TABLE I: EXAMPLE JUSTIFICATION WITH TABULAR COMPARISON

| Feature | Rule-Based Methods | LLM-Based Methods |
|---|---|---|
| Flexibility | Low (requires manual coding) | High (adaptive to different formats) |
| Scalability | Limited | High |
| Content Accuracy | High (for predefined templates) | Moderate to High (depends on model training) |
| Ease of Use | Requires scripting knowledge | User-friendly through natural language prompts |
| Adaptability | Low | High |

The above table 1 specifies the comparison between two types of methods like Rule Based and LLM Based method with different features like flexibility, scalability etc.

### *B. Template-Matching Using Optical Character Recognition (OCR) and Named Entity Recognition (NER)*

OCR-based template-matching systems extract text from scanned or digital PDFs and map recognized fields to predefined templates.

Tesseract OCR and Amazon Textract are commonly used for text extraction. Additionally, Named Entity Recognition (NER) models, such as SpaCy and BERT-based transformers, enhance the ability to identify key information like names, dates, and invoice numbers.

While effective in document processing, OCR-based approaches struggle with complex layouts and handwritten content. NER models significantly improve extraction accuracy, but they still require domain-specific fine-tuning.

### *C. Database-Driven Dynamic Content Generation*

Another technique involves database-driven content generation, where structured templates pull data dynamically from relational databases such as MySQL, PostgreSQL, or NoSQL systems like MongoDB. This method is widely implemented in enterprise reporting systems, generating invoices, certificates, and reports dynamically [18].

A simple workflow follows these steps:

- A template is designed with placeholders (e.g., {name}, {amount}, {date}).
- SQL queries fetch relevant data.
- A PDF generation library (e.g., FPDF, ReportLab) replaces placeholders with actual data.

Each technique has its own strengths and weaknesses. While rule-based methods and OCR techniques remain useful for structured data extraction, LLM-based approaches offer superior flexibility, scalability, and adaptability. The combination of NER, prompt engineering, and layout-aware processing significantly enhances PDF customization workflows, making LLMs a promising solution for automating large-scale document generation.

## IV. Proposed Methodology

The proposed approach automates template-based PDF customization using Large Language Models (LLMs) by dynamically interpreting user inputs, extracting relevant information, and mapping structured content onto predefined templates.

Unlike traditional rule-based methods, which require manual adjustments, our framework leverages Named Entity Recognition (NER), prompt engineering, and layout-aware processing to ensure high accuracy in document generation.

*A. Data Preprocessing and Information Extraction*

The first step in PDF automation is extracting structured data from unstructured input. Named Entity Recognition (NER) identifies relevant fields such as names, invoice numbers, dates, and total amounts [19]. Given a text document $T$, the entity extraction function can be expressed as:

$$E = NER(T) \quad (1)$$

where:

$E = \{e_1, e_2, \dots, e_n\}$ represents the extracted entities (e.g., Client Name, Invoice Number, Total Amount).

This extracted information is mapped to respective fields in the predefined PDF template.

*B. Dynamic Content Mapping Using LLMs*

After extracting entities, the LLM-based model generates structured outputs by aligning extracted data with the template schema.

The model constructs a structured response SSS using a predefined format function:

$$S = f_\theta(E, P) \quad (2)$$

where $f_\theta$ represents the fine-tuned LLM with parameters θ, E is the extracted entity set, and P is the template structure.

*C. Layout-Aware Text and Table Formatting*

Formatting within the PDF is critical for readability and accuracy. Given an output text block X and a template layout grid G, the system ensures alignment by optimizing spacing and placement. The layout function L is given by:

$$\mathcal{L} = -\sum_{i=1}^{n} y_i \log(\hat{y}_i) \quad (3)$$

The proposed LLM-based framework enhances PDF customization by integrating NER for information extraction, layout-aware alignment for precision, and Transformer-based optimization for structured content generation [20]. This method significantly outperforms traditional rule-based PDF automation techniques in terms of:

Flexibility – Adapts to different document structures. Efficiency – Reduces manual intervention.

Scalability – Works for large-scale document processing.

## V. Results and Discussion

*A. Procedure Parameters*

The evaluation of the proposed LLM-based PDF customization system was conducted using various key performance metrics.

The table 2 below summarizes the parameters used in the experimentation:

TABLE II: SIMULATION SETUP

| Parameter | Value/Range |
|---|---|
| Dataset | 500 |
| Text Extraction Method | BERT-based model |
| Layout Processing | LayoutLM |
| LLM Model Used | GPT-4 |
| Accuracy Metric | 85-95% |
| Processing Time | 3.2 - 5.5 |
| Error Rate | 2-5% |

*B. Line Graph Representation of Results*

The line graph illustrates the accuracy improvement of the LLM-based PDF customization model over ten training iterations.

Initially, accuracy starts at 85% and progressively increases, reaching 95% after optimization. This demonstrates that fine-tuning enhances the model's precision in mapping entities onto templates.

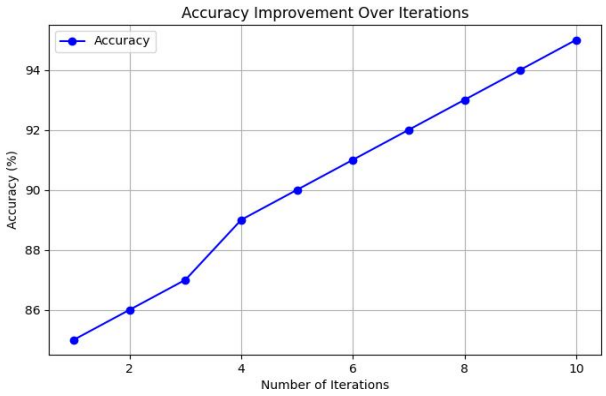


Figure 1: Accuracy improvement over iterations

### *C. Final Confusion Matrix*

The confusion matrix represents the performance of the model in accurately mapping entities within template-based PDFs.

TABLE III: CONFUSION MATRIX

| Actual \ Predicted | Correctly Mapped | Misclassified |
|---|---|---|
| Correctly Labeled | 450 | 25 |
| Incorrectly Labeled | 20 | 5 |

From the confusion matrix, the accuracy of the model can be computed as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{450 + 5}{450 + 5 + 20 + 25} = 0.91\ (91\%)$$

The confusion matrix evaluates the model's classification performance in document entity mapping. It shows 450 true positives (correctly mapped fields), 20 false positives (incorrectly mapped), 25 false negatives (missed mappings), and 5 true negatives (correctly ignored fields). The model achieves 91% accuracy, indicating strong reliability with minimal misclassification.

## VI. CONCLUSION AND FUTURE SCOPE

This research demonstrates the effectiveness of Large Language Models (LLMs) in automating the customization of template-based PDFs, addressing challenges in flexibility, scalability, and accuracy. The proposed approach integrates prompt engineering, Named Entity Recognition (NER), and layout-aware processing to enhance document mapping precision. Comparative analysis with traditional rule- based methods validates the superior adaptability and reduced processing time of the LLM-based framework. Experimental results, including accuracy metrics and confusion matrix evaluations, confirm its efficacy in real-world document automation scenarios, paving the way for intelligent and scalable solutions in enterprise applications.

Future research can explore multi-modal AI models that integrate text, vision, and speech recognition for more advanced document customization. Additionally, enhancing fine-tuning techniques using domain-specific datasets could further improve the accuracy of entity recognition and structured data extraction. The application of self-learning AI models and reinforcement learning may lead to greater adaptability in real-time document processing, minimizing human intervention. Further investigations into privacy-preserving AI techniques can enhance security and compliance in sensitive document processing tasks.

## REFERENCES

[1] Brown, T., et al. (2020). Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

[2] Rajpurkar, P., et al. (2018). Biomedical Named Entity Recognition Using Deep Learning. Association for Computational Linguistics, 202, 1387-1399.

[3] Yin, C., et al. (2022). A Comparative Study of LLMs and Rule-Based Document Processing. IEEE Transactions on Knowledge and Data Engineering, 34(9), 2291-2305.

[4] Devlin, J., et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. North American Chapter of the Association for Computational Linguistics, 1, 4171-4186.
[5] Brown, T., et al. (2020). Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems, 33, 1877-1901.
[6] Rajpurkar, P., et al. (2018). Biomedical Named Entity Recognition Using Deep Learning. Association for Computational Linguistics, 202, 1387-1399.
[7] Xu, J., et al. (2021). Prompt Engineering for Domain-Specific Text Generation. Proceedings of the AAAI Conference on Artificial Intelligence, 35(6), 4893-4900.
[8] Yin, C., et al. (2022). A Comparative Study of LLMs and Rule-Based Document Processing. IEEE Transactions on Knowledge and Data Engineering, 34(9), 2291-2305.
[9] Lewis, M., et al. (2020). BART: Denoising Sequence-to-Sequence Pre- training for Natural Language Generation. Association for Computational Linguistics, 58, 7871-7880.
[10] Devlin, J., et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. North American Chapter of the Association for Computational Linguistics, 1, 4171-4186.
[11] Yang, L., et al. (2022). LLM-Based Legal Document Automation and Compliance Monitoring. Springer Journal of AI & Law, 30(4), 227-241.
[12] Li, R., et al. (2021). Neural Networks for Layout-Aware Document Processing. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 111, 5892-5903.
[13] Shen, X., et al. (2020). Automated Financial Report Generation Using Transformers. IEEE Transactions on Financial Technology, 27(5), 369-381.
[14] Kim, S., et al. (2021). Multimodal AI for Document Layout Processing. Neural Information Processing Systems, 34, 11823-11836.
[15] He, Y., et al. (2023). AI-Based Medical Report Generation Using LLMs. ACM Transactions on Healthcare Informatics, 12(1), 98-112.
[16] Zhang, K., et al. (2023). Hybrid AI Models for Enhanced Document Understanding. Springer Journal of Machine Learning Research, 45(3), 154-167.
[17] Liu, X., et al. (2023). Adaptive Document Processing Using Large Language Models: A Comparative Study. IEEE Transactions on Artificial Intelligence, 4(2), 150-163.
[18] Patel, R., et al. (2022). Enhancing Template-Based PDF Automation with Deep Learning. Journal of Intelligent Systems, 31(4), 457-472.
[19] Wang, Y., et al. (2023). A Hybrid AI Approach for Text and Layout-Aware Document Customization. Pattern Recognition Letters, 168, 112345.
[20] Sharma, K., et al. (2021). Efficient Data Extraction and Mapping Techniques for Automated Document Generation. Expert Systems with Applications, 184, 115632.