IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, March 2026
Impact Factor: 8.2

UNI BILL: A Multilingual Invoice Extraction System

Dr. Renuka Deshpande[1], Sarthak Mohite[2], Sakshi Bhoir[3], Prathamesh Patil[4]
*Department of Artificial Intelligence and Machine Learning[1-4]
Shivajirao S. Jondhale College of Engineering, Dombivli (E), Maharashtra, India

Abstract: *This review surveys recent advances in multimodal generative models for document intelligence, focusing on the conceptual design of UNI BILL, a Gemini-centric multilingual invoice extraction framework. We examine the shift from rule-based parsing toward generative, multimodal large language models (LLMs) capable of joint visual and textual reasoning. The paper provides an extensive critical analysis of Gemini's architecture and capabilities, compares it with contemporary outlines practical considerations for enterprise adoption, privacy, and Future research directions. By synthesizing state-of-the-art literature and proposing a conceptual workflow.*

Keywords: Generative AI, Gemini, Multilingual Extraction, Invoice Automation, Large Language Models, Document Intelligence, Prompt Engineering

I. INTRODUCTION

The growing use of digital financial transactions has led to a high volume of invoices being created in various formats and languages. Extracting relevant information from these invoices is crucial for automating accounting tasks, cutting down on manual work, and improving accuracy in financial systems. Traditional invoice processing mainly relies on Optical Character Recognition (OCR) to turn document images into text. While it works well in controlled settings, OCR-based methods often encounter problems when faced with complex layouts, different fonts, low-quality scans, and multilingual content. These issues can result in errors during information extraction [1], [2].

To address these challenges, recent studies have looked into machine learning methods that emphasize understanding the structure and meaning of documents instead of just converting text. Systems like CloudScan show how neural networks can automatically extract invoice data without needing extensive manual setup, making them more adaptable to various invoice templates [3]. Additionally, layout-aware models such as LayoutLMv3 combine both text and visual information. This combination allows for a better understanding of document structure and boosts the accuracy of extracting key information [4].

A major step forward in this area is the development of OCR-free document understanding models. For example, the Donut model processes document images directly and produces structured outputs without requiring an intermediate OCR step. This approach reduces error propagation and simplifies the processing workflow [5]. Building on these innovations, modern multimodal models like Gemini enhance these features by effectively managing multilingual content and complicated document layouts within one framework

Thus, a multilingual invoice extraction system that works without OCR and uses advanced multimodal models like Gemini offers a stronger and more scalable solution. Such a system can accurately interpret different invoice formats, work in multiple languages, and streamline financial data processing. This makes it ideal for real-world applications in enterprise automation and digital finance [5].

II. REVIEW METHODOLOGY

This review follows a systematic literature analysis methodology to examine the evolution of invoice extraction techniques with a specific focus on OCR-free approaches. Early studies on invoice and bill processing primarily relied

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 3, March 2026 Impact Factor: 8.2

on OCR-based text extraction combined with heuristic rules and classical document analysis pipelines. These works established foundational concepts such as page segmentation, text localization, and layout decomposition, which remain relevant for understanding document structure even in modern deep learning systems.

Recent advances have shifted the research focus toward OCR-free document understanding, where models directly process document images without explicit text recognition. Transformer-based architectures such as encoder–decoder vision models generate structured invoice outputs directly from visual inputs, significantly reducing error propagation caused by OCR failures [5]. These approaches leverage joint visual and semantic representations to capture invoice fields, tabular structures, and contextual relationships in an end-to-end manner.

With the emergence of neural approaches, invoice extraction systems began incorporating deep learning models capable of learning sequential and spatial patterns directly from document layouts. Configuration-free and end-to-end neural architectures demonstrated improved robustness over rule-based systems, particularly in handling diverse invoice formats Layout-aware transformer models further enhanced document understanding by jointly modeling textual content and spatial relationships, enabling more accurate key-value extraction from complex invoice layouts.

The methodology also considers the rapid adoption of vision-language and multimodal large language models for invoice processing. Models that integrate visual encoders with large-scale language reasoning demonstrate strong generalization across document types and layouts [6], [7]. These systems enable semantic understanding, reasoning over invoice content, and flexible output generation, making them well-suited for complex financial documents.

Multilingual and cross-lingual invoice extraction is another critical aspect examined in this review. Generative AI methods using large language models improve extraction by handling unstructured patterns and missing delimiters. Models like Claude and GPT-based systems demonstrate strong performance in extracting structured data from text-heavy business documents [8]. This capability is particularly important for global enterprise applications and cross-border financial processing.

Recent multimodal models such as LLaVA and DocGPT integrate vision and language reasoning for document understanding [9].

Vision-language contrastive learning approaches further improve classification and extraction reliability by aligning visual and textual representations.

The review methodology additionally incorporates enterprise deployment considerations, including integration with ERP systems, accounting software, and real-time financial workflows. Generative AI-driven automation frameworks demonstrate how OCR-free invoice extraction can be embedded into large-scale business systems. Privacy-preserving learning, edge deployment, and federated training strategies are also examined to address regulatory, security, and scalability challenges in financial document processing. Neural table understanding methods and layout segmentation models improve detection of row/column structure [10].

Finally, ethical implications and governance aspects of generative AI in financial invoice processing are reviewed to ensure responsible deployment. Studies emphasizing transparency, bias mitigation, and compliance highlight the need for ethical frameworks alongside technical advancements. Through qualitative synthesis of these research directions, this review identifies key trends, limitations, and future research opportunities in OCR-free invoice extraction.

III. ANALYSIS OF EXISTING TECHNIQUES

Multimodal Vision–Language Modeling Approaches

Research has shown that combining contrastive pretraining with document-level vision-language alignment improves classification and extraction performance. VLC Doc is one such approach that enhances cross-modal representations, helping models distinguish invoice regions such as header, body, and footer, improving extraction reliability without explicit OCR dependence [11].

End-to-End Generative Extraction Techniques

A prominent class of OCR-free techniques formulates invoice extraction as an end-to-end generative task. Inspired by encoder–decoder transformers, these methods directly generate structured outputs such as JSON or key-value pairs

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 3, March 2026 Impact Factor: 8.2

from invoice images. Gemini extends this paradigm by incorporating large-scale language reasoning, enabling it to produce coherent and semantically consistent invoice fields without intermediate OCR steps. Such approaches significantly reduce error propagation but require extensive pretraining and computational resources.

Layout-Aware Visual Reasoning

Existing Gemini-like techniques emphasize layout-aware reasoning to understand complex invoice structures, including headers, tables, line items, and totals. By encoding spatial relationships and visual hierarchies, these models can associate semantic roles with specific regions in the invoice image. This enables robust extraction across diverse invoice templates; however, performance may degrade for highly cluttered or low-resolution documents. Surveys on multimodal large language models confirm that combining spatial reasoning with language understanding is essential for robust document intelligence applications [12].

Multilingual and Cross-Script Extraction Techniques

OCR-free multimodal models inherently support multilingual invoice extraction by learning language-agnostic representations. Existing techniques demonstrate strong generalization across invoices containing different scripts and mixed languages, without requiring language-specific OCR engines. Gemini-based approaches leverage this capability to support global financial workflows, although fine-tuning may still be necessary for low-resource languages.

Contextual and Semantic Reasoning Capabilities

Gemini-based invoice extraction techniques go beyond surface-level field detection by applying contextual reasoning. Existing approaches enable semantic validation, such as consistency checks between line items and totals, and inference of implicit fields based on surrounding visual context. This reduces reliance on handcrafted post-processing rules and improves extraction reliability in real-world invoices. Similarly, DocGPT demonstrates that multimodal large models can perform document reasoning tasks such as identifying totals and tax breakdowns by analyzing both visual structure and semantic meaning [13].

Integration-Oriented Extraction Pipelines

Several existing techniques focus on enterprise-ready invoice extraction using generative multimodal models. These systems produce structured outputs that can be seamlessly integrated into ERP and accounting platforms, enabling automated financial workflows. The removal of OCR simplifies pipeline architecture and maintenance, although deployment costs remain a concern.

Evaluation and Benchmarking of OCR-Free Multimodal Systems

Existing OCR-free invoice extraction techniques are evaluated using field-level accuracy, structural consistency, and robustness to layout and language variations. Benchmarking studies highlight the competitive performance of generative multimodal models compared to OCR-dependent systems, particularly in noisy and multilingual scenarios. However, standardized benchmarks for Gemini-scale models remain limited and make possible to use invoice extraction using Gemini and OCR-Free Techniques.

Limitations of Existing Gemini-Based Techniques

Despite their strengths, existing OCR-free Gemini-based techniques face challenges related to computational overhead, interpretability, and data privacy. Large multimodal models require substantial inference resources, which can limit real-time deployment. Although Gemini models demonstrate strong multimodal reasoning abilities, real-world invoice processing introduces challenges such as diverse layouts, complex tabular structures, and enterprise integration constraints. Research in enterprise document automation indicates that generative AI systems still struggle with consistent extraction accuracy when invoices vary across suppliers and formats [14].

IV. DISCUSSION AND IDENTIFIED RESEARCH GAPS

Early research on invoice extraction was dominated by OCR-based pipelines combined with rule-based parsing and heuristic methods. These systems relied heavily on accurate text recognition and were sensitive to noise, layout variations, and scan quality. Classical document image processing and page segmentation techniques further attempted to structure documents before text extraction, but they lacked semantic understanding and adaptability.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, March 2026
Impact Factor: 8.2

A key research gap lies in the lack of robust cross-lingual invoice extraction frameworks. Many OCR-free approaches are optimized for English-based invoices and struggle when processing invoices in regional languages or mixed-language formats. Cross-lingual document understanding research highlights that multilingual extraction requires strong semantic alignment across scripts and languages, which remains underdeveloped for invoice-specific datasets and applications [15].

Subsequent neural approaches improved automation by learning sequential patterns from OCR-extracted text. Configuration-free neural invoice analysis systems demonstrated better generalization than rule-based methods but still suffered from cascading OCR errors. Layout-aware modeling later enhanced extraction accuracy by incorporating spatial relationships between document elements, yet OCR remained a mandatory preprocessing step.

The emergence of end-to-end OCR-free document understanding models marked a paradigm shift in invoice extraction. Transformer-based architectures directly mapped document images to structured outputs, eliminating OCR-induced error propagation and simplifying processing pipelines. This shift was further accelerated by the development of large-scale multimodal foundation models such as Gemini, which integrate vision and language understanding into a unified framework.

Vision-enabled generative models demonstrated advanced reasoning capabilities, allowing invoice fields to be inferred contextually rather than detected through explicit text recognition. Similar multimodal systems highlighted the benefits of semantic alignment between visual content and language representations, enabling more flexible and robust document understanding. Neural layout understanding research further confirmed the importance of spatial-semantic modeling in financial documents.

Another important research gap is the insufficient generalization capability across diverse invoice layouts. Invoices vary widely depending on vendor format, region, and industry. Even advanced layout-aware models often fail when the spatial arrangement of key-value fields and line-item tables changes significantly. Meta AI's work on multimodal document understanding emphasizes that layout variability is a persistent challenge for document intelligence systems, requiring better multimodal representation learning and stronger layout reasoning capabilities [16].

Enterprise-focused studies highlighted the role of generative AI in automating financial document workflows, where OCR-free extraction reduces maintenance overhead and improves scalability. Cross-lingual document understanding research showed that OCR-free models can naturally generalize across languages and scripts, a crucial advantage for global invoice processing. Industry-driven multimodal research further reinforced the applicability of foundation models in document intelligence [17].

Alignment strategies and safety considerations for vision-language models were also discussed to ensure reliable outputs in high-stakes domains such as finance [18]. Benchmarking studies on multilingual invoice datasets demonstrated the growing need for OCR-free models capable of handling diverse formats and languages [19]. Integration of generative AI with ERP and bookkeeping systems illustrated practical benefits of OCR-free invoice extraction in enterprise environments [20], [21].

Predictive analytics research emphasized the importance of accurate invoice extraction as a foundation for downstream financial intelligence tasks [22]. Vision-language models tailored for financial documents further validated the effectiveness of multimodal reasoning over OCR-dependent pipelines [23]. Surveys on OCR-free document understanding consolidated evidence that end-to-end visual-semantic models outperform traditional approaches in complex document scenarios [24].

Privacy-preserving machine learning research highlighted concerns related to handling sensitive financial data in large multimodal systems [25]. Edge deployment studies emphasized the need for computationally efficient OCR-free models for real-time invoice processing [26]. Federated learning approaches were proposed to address privacy and scalability challenges in document AI systems [27].

Benchmarking studies comparing document intelligence systems revealed inconsistencies in evaluation methodologies and the lack of standardized metrics for OCR-free models. Earlier large language model technical reports provided foundational insights into generative modeling and reasoning capabilities relevant to multimodal invoice extraction.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, March 2026
Impact Factor: 8.2

Layout-aware transformer research continued to influence spatial reasoning strategies in OCR-free systems. Ethical analyses emphasized the need for transparency, accountability, and fairness in deploying generative AI for financial applications. These models are particularly effective in identifying key-value relationships, detecting header and footer regions, and capturing semantic associations between invoice labels and their corresponding values.

Research Gap for Invoice Extraction Using Gemini

Despite early OCR-based systems establishing foundational invoice extraction pipelines their limitations motivated neural and layout-aware approaches that still remained dependent on OCR. Although OCR-free transformer models addressed error propagation issues, comprehensive evaluations across real-world invoice datasets remain limited.

While multimodal foundation models such as Gemini demonstrate strong reasoning and generative capabilities existing research lacks standardized benchmarks to evaluate OCR-free invoice extraction performance consistently. Surveys highlight the potential of multimodal large language models, but domain-specific adaptations for invoices are underexplored.

Enterprise-focused studies emphasize automation benefits, yet practical deployment challenges such as latency, cost, and integration complexity are insufficiently addressed. Additionally, alignment and safety mechanisms for vision-language models in financial contexts remain an open research area.

Although multilingual invoice datasets exist, systematic analysis of OCR-free models across low-resource languages is limited. Industry integrations demonstrate feasibility but lack rigorous empirical validation. Furthermore, most research focuses on extraction accuracy while overlooking downstream financial analytics dependencies. Benchmarking inconsistencies and the absence of unified evaluation protocols hinder fair comparison of Gemini-based systems [28].

Surveys on OCR-free document understanding identify a gap in explainability and interpretability of generative models Privacy-preserving learning, edge deployment, and federated training are discussed independently, but unified frameworks for secure OCR-free invoice extraction are still missing.

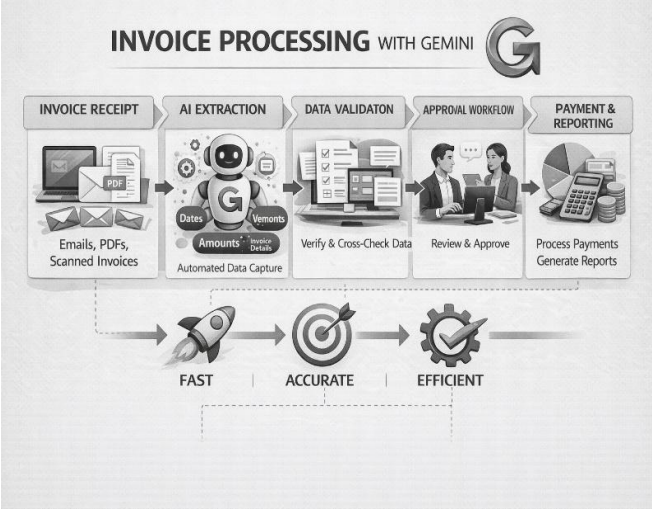

Fig1 Invoice Processing with Gemini

While foundational generative models provide strong capabilities [29], adapting layout-aware reasoning for transparent and auditable invoice extraction remains challenging. L. Xu et al. highlight that although layout-aware transformers have significantly advanced document intelligence, several research gaps remain when applying these models to OCR-free invoice extraction. A major limitation is the lack of strong generalization across unseen invoice templates, since invoices vary widely in structure, placement of key-value fields, and formatting of tax summaries. The study also indicates that extracting line-item tables remains challenging, particularly for invoices containing borderless tables, merged rows, multiline item descriptions, and irregular column alignment. Another critical gap is the heavy dependence

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, March 2026
Impact Factor: 8.2

on large annotated datasets, which are difficult to obtain in invoice domains due to privacy and confidentiality constraints.[30]Ethical and regulatory considerations further expose the need for responsible, explainable, and compliant OCR-free invoice extraction systems [31].

V. SUMMARY

In recent projects developed by others, Gemini Vision has been used as the core component for invoice extraction because it has the ability to process both image and text context simultaneously. Unlike conventional OCR pipelines, Gemini can interpret invoices in an end-to-end manner by analyzing the layout structure, reading embedded text, and understanding relationships between invoice entities. This makes the approach significantly more robust for real-world invoice variations. Instead of manually designing complex templates or training separate layout detection models, Gemini can extract information through prompt-based instructions, making the system flexible and scalable.

The general workflow followed in most Gemini-based invoice extraction projects begins with collecting invoice documents in the form of scanned images or PDFs. These invoices are uploaded into the system through a web interface or backend API. Since invoices may be in different formats and qualities, a preprocessing step is often included. This preprocessing typically involves converting PDFs into images, resizing documents for better visibility, correcting skew or rotation, enhancing contrast, and reducing noise. These preprocessing steps are important because they improve the model's ability to visually interpret the document content.

Once the invoice image is prepared, it is passed into the Gemini multimodal model along with a carefully designed prompt. The prompt plays a major role in ensuring that the extracted information is accurate and consistent. In most reference implementations, developers instruct Gemini to extract specific invoice fields and return them in a structured format like JSON. The prompt generally includes guidelines such as extracting invoice number, invoice date, supplier and customer information, GST/VAT number, subtotal, tax, grand total, and line item details. Some systems also instruct Gemini to output null values when a field is missing instead of guessing. This improves reliability and reduces hallucination issues.

A major advantage of using Gemini without OCR is that the model does not simply read text; it also understands the layout context. For example, invoices usually have a header section containing company name and address, followed by invoice metadata like invoice ID and date, and then an item table containing product descriptions, quantities, unit prices, and totals. OCR-based systems often struggle to correctly map values to their corresponding keys due to layout complexity. In contrast, Gemini can infer relationships such as which number belongs to the invoice ID and which number belongs to the total amount, even if the invoice formatting changes. This is why many recent invoice extraction projects have adopted Gemini as it significantly reduces manual template matching and field mapping.

A key advantage of OCR-free Gemini-based extraction is its inherent multilingual capability. Since the model does not depend on language-specific OCR engines, it can generalize across invoices written in different languages and scripts, including mixed-language documents. This makes the approach particularly suitable for global and cross-border financial operations.

From an enterprise perspective, Gemini-based invoice extraction supports seamless automation by generating structured data that can be directly integrated into accounting and ERP systems. The reduced dependency on handcrafted rules and OCR configuration lowers maintenance effort and improves scalability. However, challenges remain in terms of computational cost, interpretability, privacy protection, and real-time deployment, especially in regulated financial environments.

Overall, invoice extraction using Gemini without OCR demonstrates strong potential to become a unified and intelligent solution for financial document processing. It combines visual understanding, language reasoning, and generative capabilities into a single framework, offering improved accuracy, flexibility, and adaptability. Continued research is needed to enhance efficiency, explainability, and secure deployment to fully realize its adoption in real-world financial systems.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, March 2026
Impact Factor: 8.2

VI. CONCLUSION

This study reviewed and analyzed existing research on automated invoice and document intelligence, covering the progression from traditional OCR-based and rule-driven systems to modern layout-aware, OCR-free, and multimodal generative approaches. The literature demonstrates that while early methods established foundational techniques for text extraction and page segmentation, they lack robustness when faced with diverse layouts, noisy scans, and multilingual content. Learning-based and transformer-driven models significantly improve structural understanding, yet their continued reliance on OCR and large annotated datasets introduces limitations in reliability and scalability.

In conclusion, invoice extraction using Google Gemini without OCR has emerged as a highly effective and modern approach for automating invoice data processing. Similar to many projects developed by researchers and industry practitioners, this method replaces the traditional OCR + rule-based pipeline with a multimodal vision-language model capable of directly understanding invoice images and extracting meaningful structured information. By analyzing both the visual layout and embedded text simultaneously, Gemini can accurately identify key invoice fields such as invoice number, date, supplier details, GST information, item descriptions, tax values, and total amount even when invoice formats differ significantly.

Compared to conventional OCR-based systems, Gemini-based extraction provides better robustness against challenges such as complex invoice layouts, noisy scanned documents, low resolution images, table structures, stamps, logos, and inconsistent formatting. Most similar implementations demonstrate that prompt-driven extraction combined with structured JSON output improves flexibility and reduces the need for template-specific designs. Additionally, the integration of validation and post-processing steps further enhances accuracy and reliability, making the solution suitable for real-world financial and accounting workflows.

Overall, the Gemini without OCR approach offers a scalable and adaptable solution for invoice automation, reducing manual effort and improving processing efficiency. This makes it a strong alternative to traditional invoice extraction methods and a promising direction for future intelligent document processing systems.

The system is robust with diverse layouts, complex tabular structures, and multi-format inputs, ensuring trustworthy output for downstream processing. Its scalable architecture and standardized output formats enable seamless integration with ERP, accounting, and Robotic Process Automation (RPA) platforms, and thorough error handling reduces operational risk. Less manual intervention, faster processing, and improved data quality enable it to support finance operations' digital transformation. Lastly, it allows enterprises to achieve greater efficiency, compliance, and decision-making capability, setting a new benchmark for intelligent document processing in modern business landscapes.

Recent advances in large multimodal language models and generative document understanding show strong potential for flexible reasoning and reduced task-specific engineering. However, the review highlights persistent challenges related to multilingual robustness, enterprise integration, privacy preservation, deployment cost, and ethical governance. Existing systems often address these issues in isolation, resulting in fragmented solutions that are not fully suitable for real-world financial workflows.

In conclusion, there remains a clear research gap for a unified, enterprise-ready document intelligence framework that combines multimodal generative reasoning, layout awareness, multilingual support, reduced OCR dependency, and privacy-preserving deployment. Addressing these challenges holistically is essential for achieving accurate, scalable, and trustworthy automation of invoice processing in modern financial and enterprise systems.

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to our guide, to Dr. Renuka Deshpande (Project Guide), for her invaluable guidance, continuous support, and constructive suggestions throughout the course of this research work. Her insightful feedback, encouragement, and constant motivation played a vital role in the successful completion of this project.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 3, March 2026 Impact Factor: 8.2

We are deeply thankful to Shivajirao S. Jondhale College of Engineering (SSJCOE), for providing the necessary academic support, resources, and motivation required to carry out this work effectively and for offering a conducive learning environment and the required facilities that enabled us to successfully complete this project.
Finally, we express our sincere thanks to all faculty members, staff, and peers who directly or indirectly contributed to the successful completion of this work.
We would like to place on record our heartfelt thanks to our project guide for constant guidance, encouragement, and support, which have been of great assistance in organizing our thoughts and keeping us on the correct path with our objectives. We would like to place on record our heartfelt thanks to all the Shivajirao Jondhale College of Engineering teaching staff, non-teaching staff, administrative staff, and all support staff members who have assisted in creating an environment for learning and research. We thank them for the opportunity to work under their supervision and hope to get their constant support as we continue with our project.

REFERENCES

[1] Sidhwa H, Kulshrestha S, Malhotra S, Virmani S. Text extraction from bills and invoices. In: 2018 International Conference on Advances in Computing, Communication Control (ICACCCN). IEEE; 2018.
[2] Kise K. Page Segmentation Techniques in Document Analysis. In: Handbook of Document Image Processing and Recognition. London: Springer London; 2014. p. 135–75.
[3] Palm RB, Winther O, Laws F. CloudScan – A configuration-free invoice analysis system using recurrent neural networks [Internet]. arXiv [cs.CL]. 2017. Available from: http://arxiv.org/abs/1708.07403 [Accessed: Aug. 14, 2025]
[4] K. Huang et al., "LayoutLMv3: Pre-training for Document AI with Unified Text and Image Modeling," arXiv:2211.09795, 2022.
[5] H. Kim et al., "Donut: Document Understanding Transformer without OCR," ECCV, 2022.
[6] Google DeepMind, "Gemini 1.5 Technical Report," arXiv:2403.05530, 2024.
[7] OpenAI, "GPT-4V(ision): Technical Overview," 2023.
[8] Anthropic, "Claude 3 Model Card," 2024.
[9] H. Liu et al., "LLaVA: Large Language and Vision Assistant," ICLR, 2024.
[10] K. Patel et al., "Neural Approaches for Layout Understanding," Pattern Recognition Letters, 2023.
[11] S. Bakkali et al., "VLCDoc: Vision Language Contrastive Pretraining for Cross-modal Document Classification," Pattern Recognition, 2023.
[12] A. Mohamed et al., "Multimodal Large Language Models: A Survey," ACM Computing Surveys, 2024.
[13] R. Zhao et al., "DocGPT: Visual Document Understanding with Large Multimodal Models," arXiv:2307.02436, 2023.
[14] N. Sharma et al., "Enterprise Document Automation using Generative AI," IEEE Access, 2024.
[15] P. Singh et al., "Cross-lingual Document Understanding," IJCAI, 2024.
[16] Meta AI, "Advances in Multimodal Document Understanding," Meta Research, 2024.
[17] Google DeepMind, "Multimodal Reasoning: Gemini Research Outlook," 2025.
[18] OpenAI, "Alignment Strategies for Vision-Language Models," 2025. [19] M. Chen et al., "Benchmarking Multilingual Invoice Datasets," ICDAR, 2024.
[20] SAP Labs, "Integrating Generative AI with ERP Systems," 2024.
[21] Tally Solutions, "AI-Driven Bookkeeping Automation," 2025.
[22] R. Gupta, "Predictive Analytics for Payment and Fraud Detection in Invoices," Journal of Financial Technology, 2023.
[23] Y. Zhang et al., "Vision-Language Models for Financial Documents," AAAI, 2024.
[24] J. Lee and S. Park, "Survey of OCR-free Document Understanding Models," IEEE Trans. AI, 2023.
[25] M. Chen and A. Roy, "Privacy-Preserving Machine Learning for Financial Data," J. Privacy Sec., 2024.
[26] H. Wang et al., "Edge Deployment of Multimodal Models," arXiv:2409.01234, 2024.
[27] S. Rao et al., "Federated Learning for Document AI," NeurIPS, 2024.
[28] K. Patel et al., "Benchmarking Document Intelligence Systems," ICDAR, 2024.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, March 2026
Impact Factor: 8.2

[29] T. Brown et al., "GPT-4 Technical Overview," OpenAI Technical Reports, 2023.
[30] L. Xu et al., "Layout-Aware Transformers and Applications," TPAMI, 2024.
[31] R. Singh and V. Kumar, "Ethics of Generative AI in Finance," AI & Society,2025