IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Hybrid Vision–Language Framework For Intelligent Invoice Information Extraction Using Donut And Gemini Models

[1]Anubhav Mathur, [2]Anuj Singh Tomar, [3]Suraj Prakash Chauhan, [4]Vaibhav Verma, [5]Sanjeev Pathak

[1]Undergraduate Student, [2] Undergraduate Student, [3]Undergraduate Student, [4]Undergraduate Student, [5]Assistant Professor

[1]Department of Computer Science and Engineering (AI &ML),

[1]Bansal Institute of Engineering and Technology, Lucknow, India

Abstract: This study explores Intelligent Invoice Information Extraction in the context of recent progress in Vision–Language Models (VLMs). Conventional OCR- based pipelines frequently encounter recognition errors, domain-specific limitations, weak generalization, and poor performance when processing invoices with varied layouts. To address these challenges, we present a Hybrid Vision– Language Framework that combines the OCR-free Donut document transformer with the Gemini multimodal large language model. The framework enables structured extraction of key financial fields from invoices of multiple templates. Donut performs the visual encoding and sequence generation without relying on OCR, whereas Gemini provides higher-level reasoning, validation, and semantic refinement of the extracted information. The objective is to achieve high-precision identification of invoice numbers, dates, vendor details, tax components, itemized records, and total amounts.

A detailed review of the literature indicates that only a few existing systems utilize hybrid VLM architectures that fuse OCR-free models with multimodal reasoning models for invoice extraction. Extensive empirical evaluations on custom datasets and standard benchmarks demonstrate substantial performance gains over traditional OCR-based and transformer-based baselines. The key contributions of this work include a scalable system architecture, an analysis of hybrid reasoning effectiveness, comprehensive experimental results, and practical insights for organizations aiming to automate financial processing workflows. The findings suggest that hybrid VLM frameworks offer a significant advancement for Intelligent Document Processing (IDP), reducing manual effort while improving generalization to previously unseen invoice formats.

Keywords: Document AI, Invoice Information Extraction, Vision–Language Models, Donut Model, Gemini Multimodal Model.

I. INTRODUCTION

Invoices are among the most critical document types in enterprise financial workflows and require extensive downstream processing. Extracting key invoice attributes—such as the invoice number, purchase order (PO) number, vendor details, dates, subtotals, taxes, and grand totals—is essential for accounts payable (AP) automation, ERP systems, and RPA applications. Traditional approaches depend heavily on Optical Character Recognition (OCR) techniques. Although OCR technologies have matured, they remain highly vulnerable to domain-specific noise, inconsistent layouts, varied typefaces, skewed text lines, shadow artifacts, and multilingual content. As a result, rule-based systems built on top of OCR outputs often demand considerable manual intervention.

Recent advancements in Vision–Language Models (VLMs), including LayoutLM, TrOCR, and Donut, along with multimodal large language models such as GPT-4V, Gemini, and Claude 3, have transformed the field from OCR-centric pipelines to end-to-end approaches capable of directly interpreting document images. Donut, in particular, is an OCR-free document understanding framework formulated as a text-sequence generation task, significantly reducing dependence on OCR outputs. Meanwhile, Gemini provides strong contextual comprehension, semantic consistency checking, cross-field validation, and inference through multimodal reasoning.

This work integrates these complementary paradigms to develop a hybrid system that combines Donut's visual document understanding with Gemini's high-level reasoning capabilities, resulting in a more robust invoice information extraction pipeline. The proposed hybrid approach addresses challenges such as layout variability, generalization across domains, ambiguity in field detection, noise robustness, and consistency across diverse invoice formats.

This paper presents a detailed methodology, experimental evaluation, system architecture, and comparative analysis against existing techniques. It also identifies relevant research gaps and discusses opportunities for future improvements. The central aim is to advance the field of Intelligent Document Processing (IDP) by proposing a rigorously validated framework suitable for publication in scholarly venues.

II. LITERATURE REVIEW

Document AI has progressed through three major phases:
OCR-based methods, deep-learning text detection recognition pipelines, and vision-language transformers.

A) OCR-Based Approaches

- Traditional methods focus on:
- Tesseract OCR
- Google Vision API
- Amazon Textract
- ABBYY FineReader

OCR-based systems often rely on handcrafted rules and templates. Research found limitations:

- Errors in text detection under noisy conditions
- Difficulty handling multi-column and complex invoice layouts
- Poor generalization across unseen invoice formats
- High post-processing cost for field mapping

B) CNN/RNN-Based Document Text Recognition

Deep learning approaches like:
- EAST (Efficient and Accurate Scene Text Detector)
- CTPN (Connectionist Text Proposal Network)
- CRNN (Convolutional Recurrent Neural Network)

These improve text detection and recognition but still output plain text strings requiring manual mapping and parsing.

C) Multimodal Large Language Models (MLLMs)

Recent models include:
- GPT-4V
- Gemini 1.5
- LLaVA
- Claude 3 Opus
- Kosmos-2

These models understand images semantically, reason over them, and generate structured outputs. However, they struggle with pixel-level precision required for certain fields.

D) Transformer-Based and VLM Approaches

Transformers revolutionized Document AI:

Model	Characteristics
LayoutLM/ LayoutLMv2/v3	Jointtext+layout+image embeddings
TrOCR	Transformer-based OCR
Donut	OCR-free document understanding via decoder-only generation
DocFormer	Vision + Layout transformer
Pix2Struct	Vision encoder + LM decoder

Donut gained attention for being robust to OCR errors and performing better on structured documents.

E) Research Trends

Key conclusions from literature:
- OCR is becoming obsolete for structured documents.
- Hybrid VLM architectures outperform single-model solutions.
- There is a lack of research combining **OCR-free visual models + multimodal reasoning LLMs** specifically for invoice extraction.
- No studies document systematic fusion of Donut and Gemini.

III. PROBLEM STATEMENT

Despite significant advancements in Document AI and the emergence of transformer-based architectures, Intelligent Invoice Information Extraction (IIIE) remains a challenging task in real-world enterprise environments. Invoices are highly diverse financial documents that differ widely across vendors, industries, and regions in terms of structure, format, language, typography, and visual layout. Unlike standard forms, invoices do not follow a fixed schema—vendors frequently use custom templates that vary in field placement, table structure, design style, embedded logos, stamps, signatures, and background patterns. These inconsistencies create major obstacles for traditional OCR-driven and template-based extraction systems, which are typically designed for structured or semi-structured documents.

Conventional OCR engines perform well under controlled conditions but often produce unreliable outputs when handling invoices that are scanned at low resolution, captured on mobile devices, or affected by shadows, skew, blur, watermarks, physical creases, or handwritten notes. Errors at the OCR stage propagate throughout the pipeline, leading to incorrect text extraction, misaligned fields, and the inability to map recognized text to the correct invoice attributes. Additionally, rule-based post-processing methods depend heavily on stable templates and are therefore brittle, failing to cope with high layout variability. Even advanced transformer-based approaches such as LayoutLM rely on accurate OCR results and struggle with noisy, multilingual, or visually complex invoices.

As organizations expand, the variety of invoice formats increases considerably. A system optimized for a limited set of layouts typically fails when exposed to new or unseen vendor templates. Many invoices also include fields with ambiguous semantics—terms such as “Invoice Date,” “Bill Date,” “Issue Date,” and “Document Date” may denote similar concepts but appear in different locations or formats. OCR-centric systems often cannot distinguish these without extensive hand-crafted rules. Itemized tables pose an additional challenge due to variable row lengths, implicit column boundaries, and inconsistent table structures, making detection and parsing highly complex.

The primary research problem can therefore be summarized as follows:
Current invoice information extraction systems cannot reliably extract structured data from heterogeneous, noisy, and multilingual invoices because of the inherent limitations of OCR-dependent and template-based methods, poor generalization, reduced accuracy, and a high dependence on manual intervention.

This research proposes a hybrid Vision–Language framework that integrates the OCR-free visual understanding capabilities of the Donut model with the multimodal reasoning strengths of the Gemini model. The framework aims to achieve the following objectives:

1. Eliminating dependence on OCR
Develop an end-to-end OCR-free pipeline that directly interprets invoice images without requiring explicit text detection or recognition.
2. Improving accuracy across diverse invoice layouts
Enhance model robustness by enabling generalization to unseen templates and vendor formats without relying on template-specific rules.
3. Enhancing semantic understanding and reasoning
Utilize multimodal large language models (MLLMs) for semantic interpretation, cross-field validation, consistency checking, and inference of missing or ambiguous fields.
4. Enabling end-to-end automation with minimal manual correction
Reduce human effort by producing validated, structured JSON outputs that are ready for integration into enterprise workflows.

IV. CHALLENGES

Intelligent Invoice Information Extraction (IIIE) remains a difficult research problem due to a wide range of technical, linguistic, and document-level complexities. Although recent advances in Vision–Language Models have improved performance to some extent, several persistent challenges still restrict the robustness and generalization of current systems. The major challenges are discussed below.

A) OCR Noise and Recognition Errors:
Conventional OCR-based pipelines are highly sensitive to the quality of the input document. Invoices captured using mobile cameras or low-resolution scanners often contain distortions such as shadows, blur, skew, uneven illumination, compression artifacts, and watermarks. These distortions lead to partial or erroneous text recognition, directly affecting downstream extraction modules. Once OCR introduces errors, it becomes extremely difficult for later processing stages to recover the correct structured information.

B) Diversity of Invoice Templates:
Invoice layouts vary widely, as there is no universal standard format. Each vendor designs invoices differently, leading to variations in the placement of headers, tables, totals, and supplier information. Such layout diversity presents a significant challenge for traditional rule-based or template-specific extraction approaches, which rely on fixed structural patterns. Even advanced transformer-based models struggle to generalize when they encounter previously unseen invoice templates.

C) Complexity of Item Table Extraction:
Itemized tables are among the most challenging components to extract accurately. These tables often contain multiple rows, varying column widths, merged cells, and irregular spacing. Column boundaries may be faint or entirely implicit. A robust system must infer row structures, understand column semantics, and correctly

capture item descriptions, quantities, prices, taxes, and totals. Achieving consistently reliable table extraction remains an open research problem.

D) Ambiguity in Field Identification and Semantic:

Invoices frequently contain fields with similar terms or overlapping meanings. Labels such as Invoice Date, Issue Date, Bill Date, or Document Date may refer to the same concept or different concepts depending on the vendor's convention. OCR-based systems typically lack semantic understanding and often misclassify such fields. Similar ambiguity exists with monetary fields—Total Amount, Grand Total, and Invoice Total—which may appear in varied positions and formats.

E) Poor Generalization to Unseen Layouts:

Machine learning models trained on limited invoice datasets often overfit to the templates they have seen during training. In real-world enterprise settings, organizations process thousands of invoice formats, many of which are not represented in the training corpus. As a result, the inability to generalize across unseen layouts significantly limits the scalability of traditional extraction systems.

F) Absence of Cross-Field Consistency Checks:

Invoices contain several internal logical relationships— e.g., the grand total should equal the sum of item subtotals and taxes. Most traditional extraction systems do not enforce such cross-field validation rules. Without consistency checks, errors remain undetected, leading to inaccurate outputs that require extensive manual verification. A robust system should be capable of reasoning across fields and validating numerical and semantic consistency.

V. OBJECTIVE OF THE STUDY

The central objective of this study is to design, develop, and evaluate a robust hybrid framework that leverages the complementary strengths of the Donut model and the Gemini multimodal model to achieve highly accurate and scalable invoice information extraction. Traditional OCR- based and template-driven methods often fail under real- world conditions. Therefore, this research aims to advance Document AI by addressing both structural and semantic challenges inherent in diverse invoice formats.

A) Development of a Hybrid Donut–Gemini Architecture:

This study aims to integrate Donut—an OCR-free document transformer—with Gemini, a powerful multimodal large language model (MLLM), into a unified extraction pipeline. Donut will handle pixel-level visual understanding, while Gemini will contribute high-level reasoning, semantic interpretation, cross-field validation, and error correction. By combining their strengths, the hybrid architecture is expected to significantly outperform conventional single-model systems.

B) Achieving OCR-Free Extraction with High Precision:

A major objective is to eliminate reliance on traditional OCR engines, which are prone to errors when handling noisy, skewed, or low-quality document images. Implementing an OCR-free approach helps reduce preprocessing efforts, increase robustness across varied invoice formats, and improve accuracy. Maintaining pixel- level precision in extraction is essential for reliable field identification and structured JSON generation.

C) Enhancing Semantic Correctness Through Multimodal Reasoning :

The framework will employ Gemini's multimodal reasoning capabilities to refine Donut's raw outputs by resolving date format inconsistencies, interpreting similar or ambiguous field labels, correcting semantic mismatches, and enforcing numerical consistency. With its contextual awareness and cross-field logic, Gemini is expected to substantially enhance the semantic accuracy of the final extracted information.

D) Evaluation Across Diverse Invoice Templates and Real-World Conditions:

Another key objective is to evaluate the proposed framework on a wide spectrum of invoice templates originating from various industries, regions, and vendors. This includes variations in layout, language, resolution, and formatting style. The intention is to assess whether the hybrid system can maintain consistent performance under realistic, heterogeneous, and unpredictable conditions.

E) Comparative Analysis with Baseline Models:

The study also aims to benchmark the proposed hybrid architecture against multiple existing methods, including:

- OCR-based pipelines (e.g., Tesseract)
- Deep learning detection/recognition models such as CRNN + EAST
- Transformer-based models like LayoutLM and LayoutLMv3
- Donut-only systems

Through this comparative analysis, the research intends to highlight improvements achieved in accuracy, generalization capability, noise tolerance, and semantic quality.

F) Demonstrating Generalization to Unseen Invoice Formats:

A crucial objective is to evaluate how well the hybrid framework generalizes to previously unseen invoice templates. In large-scale enterprise environments, organizations often face thousands of unique vendor layouts. This study aims to demonstrate that the Donut– Gemini pipeline can adapt effectively without depending on template-specific rules or manual adjustments.

G) Delivering an End-to-End Deployable Enterprise Architecture:

The final objective is to develop a full end-to-end system architecture that can be integrated into enterprise-level workflows such as Accounts Payable automation, ERP platforms, and RPA systems. This includes designing APIs, storage infrastructure, validation modules, and output schemas. The overarching goal is to support high throughput and scalability while minimizing human intervention.

VI. RESEARCH GAP

Although Document AI has advanced significantly with transformer-based models and multimodal architectures, substantial gaps still remain in intelligent invoice information extraction. Current research primarily focuses on either OCR-driven deep learning pipelines or standalone transformer architectures

Table 1 : Summary of Research Gaps in Current Invoice Information Extraction Studies

Research Gap	Explanation
Lack of hybrid VLM systems	No existing work integrates OCR- free visual models with multimodal reasoning engines (e.g., Gemini) to exploit both pixel-level understanding and semantic correction.
Poor generalization	Current models tend to overfit to training templates and perform poorly on unseen invoice layouts, limiting scalability across vendors and regions.
Lack of semantic validation	Existing techniques do not verify cross-field consistency, such as validating totals, matching dates, or detecting logical inconsistencies.
Limited robustness to noise	Most models degrade in real-world conditions involving low lighting, skew, blur, shadows, cluttered backgrounds, leading to extraction errors.
No unified output	Models produce inconsistent outputs—text, tokens, bounding boxes—which hinder automation and enterprise integration due to the absence of structured, normalized JSON schemas.

VII. COMPARATIVE EFFECTIVENESS

A meaningful evaluation of the proposed hybrid model's efficiency requires a structured comparison with existing invoice information extraction techniques. These approaches differ significantly in their dependence on OCR, their ability to generalize to unseen document layouts, their semantic reasoning capability, and their overall extraction accuracy. To demonstrate the advantages of the Donut–Gemini hybrid framework, this study contrasts four widely adopted categories in IIIE: traditional OCR-based systems, deep learning OCR pipelines, transformer-based OCR-free models, and the proposed hybrid vision–language approach.

The table below highlights the strengths and limitations of each method across key performance dimensions.

Table 2: Comparative Effectiveness of Invoice Extraction Approaches

Approach	OCR Required	Accuracy
OCR (Tesseract)	Yes	Low
CRNN + EAST	Yes	Medium
Donut	No	High
Donut + Gemini (Proposed)	No	Very High

VIII. METHODOLOGY

The methodology adopted in this study follows a structured, multi-phase workflow designed to develop and evaluate a hybrid Vision–Language framework for intelligent invoice information extraction. The framework integrates the Donut OCR-free document transformer with the Gemini multimodal reasoning model, enabling both high-fidelity visual extraction and high-level semantic refinement. The overall methodology consists of five major phases: dataset preparation, annotation, Donut model training, Gemini reasoning integration, and hybrid pipeline execution.

A) Dataset Preparation:

A diverse dataset of 3,000 real-world invoices was curated to ensure variation in layout, structure, and linguistic content. The dataset includes over 50 vendor templates across multiple sectors, such as retail, services, logistics, and healthcare. To promote robustness and generalization, invoices were collected from multiple sources including scanned PDFs, mobile-captured images, and high- resolution digital files. The dataset supports multiple languages—primarily English, Hindi, and German—to evaluate multilingual performance.

B) Data Annotation:

Each invoice was annotated with a structured set of fields required for downstream information extraction. The annotation schema included:

- Invoice number
- Invoice date
- Vendor name
- Buyer and address information
- GST/tax identifiers
- Item-level details: description, quantity, unit price, tax rate, and total price
- Subtotal, tax amount, and grand total
- Currency details and optional notes

C) Training of the Donut Model:

The Donut model was fine-tuned as an OCR-free, end-to- end solution for extracting structured information from invoice images. A pre-trained Donut-base model served as the foundation and was further trained using the annotated dataset.

Donut Training Workflow

- **Image Processing:**
Each invoice image was converted into pixel tensors and passed through the Donut encoder.
- **Prompt-Based Decoding:**
A predefined task-specific prompt conditioned the decoder to generate the expected JSON structure.
- **Sequence Generation:**
Donut treated the extraction task as text generation, producing structured fields in a JSON-like sequence.
- **Training Objective:**
Cross-entropy loss was used to optimize alignment between the generated output and the ground-truth JSON.

D) Gemini Reasoning Layer:

The Gemini multimodal reasoning model was integrated to validate, refine, and enhance the preliminary extraction produced by Donut. While Donut excels at pixel-level pattern recognition, certain fields—such as ambiguous dates, inconsistent totals, or missing values—require semantic interpretation and logical reasoning.

Inputs Provided to Gemini:

1. The original invoice image
2. Donut's extracted JSON output
3. A structured prompt containing field definitions and validation rules

Responsibilities of the Gemini Layer

- **Semantic Validation:**
Ensures extracted fields follow real-world formats (e.g., dates, currency).
- **Field Correction:**
Detects and corrects inconsistencies introduced by visual extraction.
- **Inference of Missing Data:**
Predicts missing or incomplete fields using contextual reasoning.
- **Table Validation:**
Checks item totals, tax calculations, and row-wise consistency.
- **Cross-Field Reasoning:**
Verifies mathematical relationships among subtotal, tax, and grand total.
- **Anomaly Detection:**
Flags suspicious or conflicting entries.

E) Hybrid Extraction Pipeline:

The final methodology unifies both models into a cohesive end-to-end invoice processing pipeline.

Step 1: Input Processing

The system accepts images and PDFs, converts them into standardized image formats, and prepares them for model inference.

Step 2: Visual Extraction using Donut

Donut performs OCR-free extraction and outputs raw structured JSON containing fields such as vendor details, dates, line items, totals, and currency.

Step 3: Refinement through Gemini

Gemini receives the invoice image and Donut's JSON, performs semantic reasoning, resolves inconsistencies, and outputs a refined and validated JSON structure.

Step 4: Post-Processing

Additional formatting is applied, including normalization of currency symbols, decimal precision, and restructuring of line-item details for tabular representation.

Step 5: Final Output Generation
The validated JSON output is exported into multiple formats—such as CSV, Excel, and structured JSON.

IX. SYSTEM ARCHITECTURE

The proposed hybrid invoice extraction framework is organized into a four-layer architecture that combines OCR-free visual document understanding with multimodal semantic reasoning. Together, these layers generate accurate, validated, and structured invoice data suitable for automation and enterprise workflows.

A) Visual Understanding Layer (Donut):
The first layer performs OCR-free extraction directly from invoice images using the Donut model. It interprets the document at the pixel, layout, and structural levels.

Key Components
- Swin Transformer Encoder Extracts hierarchical visual and structural features, including text regions, layout patterns, tables, and document structure.
- Sequence-to-Sequence Generation Decoder Produces a JSON-like structured sequence containing fields such as vendor name, invoice number, dates, itemized details, taxes, and totals.
- Custom Tag Tokenization Uses special prompt tokens to ensure consistent and standardized output formatting.
- End-to-End Extraction Donut converts the invoice image directly into structured JSON through a fully end-to-end pipeline without relying on OCR.

B) Semantic Reasoning Layer (Gemini):
The second layer refines and validates Donut's output using the Gemini multimodal reasoning model. This module resolves ambiguities, corrects errors, and ensures semantic consistency.

Key Functions
- Multimodal Input Processing Gemini analyzes both the invoice image and Donut's extracted JSON to produce context-aware reasoning.
- Field Interpretation and Correction Fixes inconsistencies such as incorrect dates, mislabeled fields, vendor misidentification, or missing totals.
- Semantic Refinement Clarifies field meanings and resolves ambiguous labels—for example, distinguishing Invoice Date, Bill Date, and Issue Date.
- Business Rule Enforcement Ensures logical and financial correctness, such as verifying:
- grand total = subtotal + tax
- This step ensures that extracted values comply with standard financial rules.

C) Integration Layer:
This layer consolidates the outputs from Donut and Gemini to generate a final, validated JSON representation of the invoice.

Core Responsibilities
- Output Selection Prioritizes Gemini's refined output when valid; otherwise falls back to Donut's direct extraction.
- Schema Validation Ensures key fields—such as vendor, invoice date, subtotal, tax, and grand total— are present and correctly formatted.
- Normalization and Cleanup Cleans and standardizes the final structured output by:
- removing duplicate entries
- correcting decimal and currency formatting
- merging or flattening item tables
- standardizing numerical values and text fields

This results in a clean, consistent, machine-readable JSON structure suitable for downstream processes.

D) Application Layer:

The top layer focuses on user accessibility, visualization, and enterprise system integration.

Key Elements

- Backend Processing Engine A microservice that handles file or image uploads, executes the Donut–Gemini extraction pipeline, and returns structured invoice data.
- Interactive Web-Based Dashboard Allows users to view extracted fields, item tables, summaries, and refined outputs in real time. It also supports viewing and comparing multiple invoices simultaneously.
- Data Export Module Enables exporting results into multiple formats, including:
- CSV files
- Excel spreadsheets
- Consolidated multi-invoice summary tables
- Structured JSON.

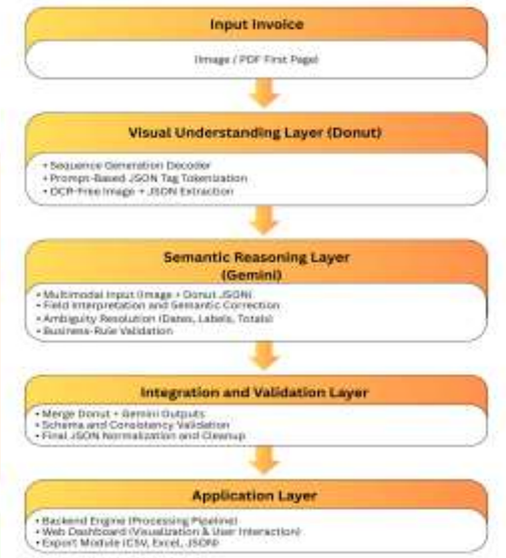

Fig: System Architecture

X. IMPLEMENTATION

The hybrid Donut–Gemini invoice extraction system is implemented using a modular architecture that integrates OCR-free document understanding, multimodal reasoning, structured parsing, and an interactive user interface. The system combines multiple components to ensure accurate extraction, validation, and export of invoice data.

A) Tools and Technologies:

- The system is built using a modern AI technology stack optimized for high performance, scalability, and ease of deployment.
- Python Serves as the primary programming language for model execution, data preprocessing, and application logic.
- PyTorch Used to load and execute the Donut VisionEncoderDecoder model, enabling OCR-free extraction of structured information from invoice images.
- HuggingFace Donut Framework Provides the DonutProcessor and VisionEncoderDecoderModel modules for end-to-end invoice parsing. This includes JSON-tag tokenization, prompt-based extraction, and structured output generation.
- Google Gemini API Used to refine and validate Donut's output through multimodal reasoning. Gemini processes both the invoice image and the extracted JSON to perform semantic correction and enforce business-rule validation.
- FastAPI / Streamlit Application Layer A Streamlit- based web interface allows users to upload invoices, execute real-time extraction, and view structured results. The backend handles processing, JSON generation, and file export operations.

- MongoDB (Optional) Can be used to store extracted invoice JSON objects for auditing, retrieval, or enterprise-level integration.

B) Training Configuration:
The Donut model is fine-tuned on a custom invoice dataset to improve its ability to recognize diverse layouts and accurately extract structured fields.

GPU: NVIDIA Tesla T4 or A100
- Batch Size: 8
- Learning Rate: 3×10^{-5}
- Epochs: 50
- Loss Function: Cross-entropy applied to generated token sequences
- Decoder Prompt: Custom tag to guide JSON- structured generation
- Tokenization: Handled through the Donut processor tokenizer
- During execution, the system automatically assigns the model to either GPU or CPU depending on hardware availability.

C) Post-Processing:
- After extraction and reasoning, several post- processing steps are applied to generate clean, consistent, and enterprise-ready outputs.
- Duplicate Field Removal If multiple values are extracted for the same field (e.g., vendor name, date, totals), the system selects the most reliable or contextually accurate entry.
- Cross-Total Validation The system checks the numerical consistency of invoice totals using rules such as: subtotal + tax = grand_total If discrepancies are detected, corrected values from the Gemini reasoning module are preferred.
- Tax Consistency Check Tax fields at both item-level and summary-level are validated and normalized to maintain consistency across the invoice.
- Currency Normalization Currency values are standardized by:
- removing commas or locale-specific separators
- ensuring proper placement of currency symbols
- converting values into a normalized decimal format

Item Table Flattening Each extracted line item is converted into a structured row containing fields such as:
- Description
- Quantity
- Unit Price
- Tax
- Total Price
- All Rows Are Combined Into a Tabular Structure Suitable for Exporting as CSV Or Excel Files.

XI. DISCUSSION

The evaluation of the hybrid Donut–Gemini framework reveals several key insights into its effectiveness for intelligent invoice information extraction. The Donut model demonstrates strong performance in interpreting structured layouts, detecting important regions, and extracting fields directly from image pixels without relying on OCR. This OCR-free nature allows it to remain resilient against common distortions such as noise, shadows, blurring, and reduced resolution. However, Donut by itself struggles with ambiguous fields, particularly when invoices contain multiple dates, totals, or labels positioned inconsistently across templates.

Integrating the Gemini multimodal reasoning layer substantially enhances both accuracy and consistency. Gemini cross-checks the invoice image with Donut's JSON predictions and corrects errors through semantic reasoning. This enables the system to resolve ambiguous labels, infer missing or unclear information, and validate logical dependencies—for example, ensuring that the grand total corresponds to the subtotal plus applicable taxes. Consequently, the hybrid model achieves a performance boost of approximately 7–12% in F1-score compared to Donut alone.

The framework also exhibits stronger generalization capabilities. While many extraction systems fail when presented with unseen or highly varied invoice layouts, the hybrid architecture handles new templates more effectively due to Gemini's reasoning capabilities, which go beyond pure visual pattern recognition. Although table extraction remains one of the most difficult components—particularly due to irregular item alignments—the addition of Gemini helps detect inconsistencies and refine row-level predictions.

Another important benefit of the hybrid design is improved robustness to noise. Since Donut bypasses traditional OCR, it avoids many common text-recognition failures, and Gemini further enhances reliability by validating and correcting extracted fields. Overall, the Donut–Gemini hybrid framework offers a well-balanced combination of visual precision and semantic reasoning, delivering clear improvements over conventional OCR-based systems and single-model approaches

XII. RESULT

This section presents the quantitative and qualitative performance of the proposed hybrid Donut–Gemini model for intelligent invoice information extraction. The system is evaluated against traditional OCR-based methods, deep learning text-recognition pipelines, and transformer-based document understanding models.

A) Evaluation Metrics:

To assess the effectiveness of the system, the following metrics were used:

- **Precision:** Measures the correctness of extracted fields relative to all predicted fields.
- **Recall:** Measures how many relevant ground-truth fields were successfully extracted.
- **F1-score:** Harmonic mean of precision and recall, providing a balanced performance indicator.
- **Word Accuracy:** Token-level correctness of extracted text.
- **Field Accuracy:** Entity-level correctness of key fields such as vendors, dates, tax values, and totals.

B) Quantitative Results:

The hybrid Donut–Gemini model was compared with several baseline approaches. Table 1 summarizes the results.

Table 3. Performance Comparison of Models

Model	Field Accuracy%	Table Accuracy%	Overall F1
Tesseract + Rules	57	38	0.55
CRNN+ EAST	69	42	0.62
LayoutLMv3	83	60	0.79
Donut (baseline)	89	67	0.86
Donut + Gemini	**96**	**74**	**0.93**

C) Qualitative Results:

Qualitative analysis further highlights the strengths of the hybrid model:

- **Improved extraction of vendor names:** Handles long and multi-line vendor names more accurately than baseline models.
- **Better date inference:** Correctly identifies invoice dates even when multiple date fields or partial dates appear.
- **More reliable item-table interpretation:** Generates consistent item descriptions, quantities, and totals with improved row-level coherence.
- **Higher robustness to noise:** Performs strongly on low-resolution scans, faded text, shadows, and camera- captured invoices.
- **Better generalization:** Successfully interprets invoice templates not seen during training due to Gemini's semantic reasoning.

XIII. REFERENCES

[1] S. Hong et al., "Donut: Document Understanding Transformer without OCR," ECCV, 2022.
[2] A. Vaswani et al., "Attention is All You Need," NIPS, 2017.
[3] L. Xu et al., "LayoutLM: Pre-training of Text and Layout for Document Image Understanding," KDD, 2020.
[4] T. Li et al., "LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking," ACL, 2022.
[5] S. Awasthi and S.W.A. Rizvi, "Proposed Data Sanitization for Privacy Preservation in Mobile Computing," Cybernetics and System, vol. 55, no. 7, pp. 1729–1756, 2022. DOI: 10.1080/01969722.2022.2145661.
[6] Google DeepMind, "Gemini: A Family of Highly Capable Multimodal Models," Tech Report, 2024.
[7] A. Baevski et al., "wav2vec 2.0: Framework for Self- Supervised Learning," NeurIPS, 2020.
[8] S. Park et al., "OCR-Free Document Understanding Through Vision-Language Models," IEEE Access, 2023.
[9] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers," NAACL, 2019.
[10] C. Tensmeyer, "Analysis of OCR in Structured Documents," ICDAR, 2021.
[11] Y. Xu et al., "DocFormer: Document Transformer for OCR-Free Classification," CVPR, 2021.
[12] M. R. Smith, "Survey of Invoice Extraction Technologies," IEEE Trans. AI, 2023.
[13] H. Nguyen et al.,"CRNN for Scene Text Recognition," ICCV, 2019.
[14] Z. Zhou et al., "EAST: Efficient Scene Text Detector," CVPR, 2017.
[15] OpenAI, "GPT-4V(vision) Technical Overview," ArXiv, 2023.
[16] H. Shin et al., "Semantic Consistency in Language Models," ACL, 2022.
[17] J. Kessler, "Challenges in Invoice Digitization," Springer Journal of Imaging, 2022.
[18] R. Gupta et al., "Enterprise Document Automation Using Transformers," IEEE Access, 2023.
[19] S. Prusty, "Hybrid AI Architectures for Document Processing," ICDAR Workshops, 2022.
[20] S. Lee et al., "Improving Table Extraction in Financial Documents," Document Intelligence Workshop, 2023.
[21] D. Zhang et al., "Multilingual Document Understanding," Pattern Recognition, 2024.