IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# Intelligent Bill Data Extraction Using Qwen VL2 Vision-Language Model: A Key-Value OCR Approach

[1]Vikrant Kamble, [2]Ankit Bhosale, [3]Jyotiraditya Chavan, [4]Harsh Parab, [5]Shreyas Gadave, [6]Prof. Swati Ambi

[1,2,3,4,5]Student, [6]Assistant Professor

[1,2,3,4,5,6]Dept. of Computer Science and Engineering

[1,2,3,4,5,6]D.Y. Patil College of Engineering and Technology, Kolhapur, India

***Abstract:*** This paper presents an intelligent Optical Character Recognition (OCR) system for automated bill data extraction using Qwen VL2, a state-of-the-art vision-language model (VLM). Unlike traditional OCR pipelines that rely on Tesseract or rule-based pattern matching, the proposed system leverages the multimodal capabilities of Qwen VL2 to process bill images and directly return structured key-value pairs corresponding to fields such as consumer number, billing amount, due date, and service period. The system integrates a Node.js backend with a MongoDB database for scalable storage and retrieval. Experimental evaluation on 200 diverse real-world electricity and utility bills demonstrates superior extraction accuracy compared to conventional OCR approaches, achieving 94.6% overall field-level accuracy versus 73.0% for Tesseract. The architecture is lightweight, extensible, and deployable in cloud or on-premise environments.

**Keywords – Qwen VL2, Vision-Language Model, OCR, Bill Processing, Key-Value Extraction, Multimodal AI, Document Automation, Node.js, MongoDB.**

## I. INTRODUCTION

The digitization of physical documents remains a persistent challenge across industries, including banking, utilities, healthcare, and government. Electricity bills, invoices, and receipts contain structured information that, when extracted automatically, can power downstream workflows such as payment processing, analytics, and auditing. According to a 2023 NASSCOM report, Indian enterprises lose an estimated 30% of productivity to manual document handling, highlighting the urgency for automated extraction systems.

Traditional Optical Character Recognition (OCR) engines such as Tesseract perform raw text extraction from images and rely on handcrafted regular expressions or template-based parsers to identify specific fields. This approach is brittle: minor variations in bill layout, font, or scan quality lead to parsing failures and require continuous maintenance of extraction rules for each utility provider.

The emergence of large vision-language models (VLMs) presents a transformative opportunity. Qwen VL2, developed by the Qwen Team at Alibaba Cloud [7], combines a high-performance visual encoder with a large language model backbone, enabling end-to-end understanding of document images without a separate OCR stage. By prompting the model with a natural language instruction to extract specific fields, the system returns structured key-value pairs directly from the image, eliminating the need for layout-specific parsing rules.

This research proposes, implements, and evaluates an intelligent bill processing system built around Qwen VL2. The contributions are: (1) a novel application of a multimodal VLM to structured bill key-value extraction; (2) a production-ready Node.js API backend with MongoDB integration; (3) a comparative evaluation against Tesseract-based extraction on 200 real-world bills; and (4) an analysis of failure modes and future improvement directions.

## II. LITERATURE SURVEY

Optical Character Recognition (OCR) has evolved significantly over the past six decades, transitioning from rule-based systems to advanced deep learning approaches. Early OCR systems relied on template matching and handcrafted feature extraction techniques, which were limited in handling variations in font, layout, and image quality [1]. The introduction of the Tesseract OCR Engine marked a major milestone, becoming one of the most widely adopted open-source OCR systems due to its flexibility and extensibility [2]. However, traditional OCR systems primarily focus on text recognition and require additional rule-based parsing for extracting structured information, making them less robust for complex document layouts.

The emergence of deep learning around 2014 significantly improved OCR performance. Convolutional Recurrent Neural Networks (CRNNs) [3] combined convolutional layers for feature extraction with recurrent layers for sequence modeling, achieving state-of-the-art results in scene text recognition tasks. Later, transformer-based architectures such as TrOCR [4] leveraged pre-trained vision and language models to further enhance recognition accuracy across both printed and handwritten text. Despite these advancements, such models often require large annotated datasets and struggle with generalization to unseen document formats.

Beyond text recognition, Key Information Extraction (KIE) has become an important research area. Models like LayoutLM [5] introduced layout-aware pre-training by incorporating spatial relationships between text elements, significantly improving extraction performance for structured documents such as invoices and forms. Similarly, Donut [6] proposed an end-to-end approach that eliminates the need for explicit OCR by directly mapping document images to structured outputs. While effective, these models often require fine-tuning on domain-specific datasets and lack flexibility in handling diverse document types without retraining.

Recently, large vision-language models (VLMs) have emerged as a powerful paradigm for document understanding. Models such as Qwen VL2 [7] integrate high-resolution visual encoders with large language models, enabling zero-shot and instruction-based extraction of structured information from images. These models overcome the limitations of traditional OCR pipelines by eliminating the need for separate text detection, recognition, and rule-based parsing stages. Furthermore, their multilingual capabilities and ability to handle complex layouts make them highly suitable for real-world applications involving diverse document formats.

However, despite these advancements, challenges remain, including computational cost, latency, and performance degradation on low-quality or highly distorted images. Existing research also lacks comprehensive evaluation of VLM-based approaches in real-world bill processing scenarios. This gap motivates the proposed system, which leverages Qwen VL2 for efficient, accurate, and scalable key-value extraction from utility bills without requiring domain-specific fine-tuning.

## III. PROBLEM STATEMENT AND OBJECTIVES

### A. Problem Statement

Utility providers across India process millions of bills monthly from multiple formats across dozens of distribution companies. Existing rule-based OCR systems fail on non-standard layouts and require expensive per-provider rule engineering. Manual data entry introduces a 5–8% error rate and is unsustainable at scale. There is a critical need for a generalizable, low-maintenance, and accurate bill field extraction system that works across provider layouts without retraining.

In addition, variations in language, font styles, and image quality further degrade the performance of traditional OCR systems. These challenges highlight the need for an intelligent solution capable of understanding document context and extracting structured information reliably across diverse real-world scenarios.

### B. Objectives

- To develop a zero-shot bill field extraction system using Qwen VL2 requiring no domain-specific fine-tuning.
- To design a scalable Node.js REST API integrating the VLM inference endpoint with MongoDB storage.
- To evaluate extraction accuracy on a benchmark of 200 real-world electricity bills from five utility providers.

• To compare VLM-based extraction against the Tesseract OCR baseline quantitatively across all target fields.
• To provide a deployable, extensible reference architecture for document extraction using modern VLMs.
• To improve robustness against variations in document layout, language, and image quality.
• To reduce manual effort and processing time by enabling automated, end-to-end document understanding.

## IV. SYSTEM ARCHITECTURE

The proposed system consists of four major components: (1) a web-based frontend for bill image upload, (2) a Node.js REST API backend, (3) the Qwen VL2 inference engine, and (4) a MongoDB database for structured storage. The overall data flow is: Image Upload → Preprocessing → VLM Inference → JSON Parsing → DB Storage → Response.

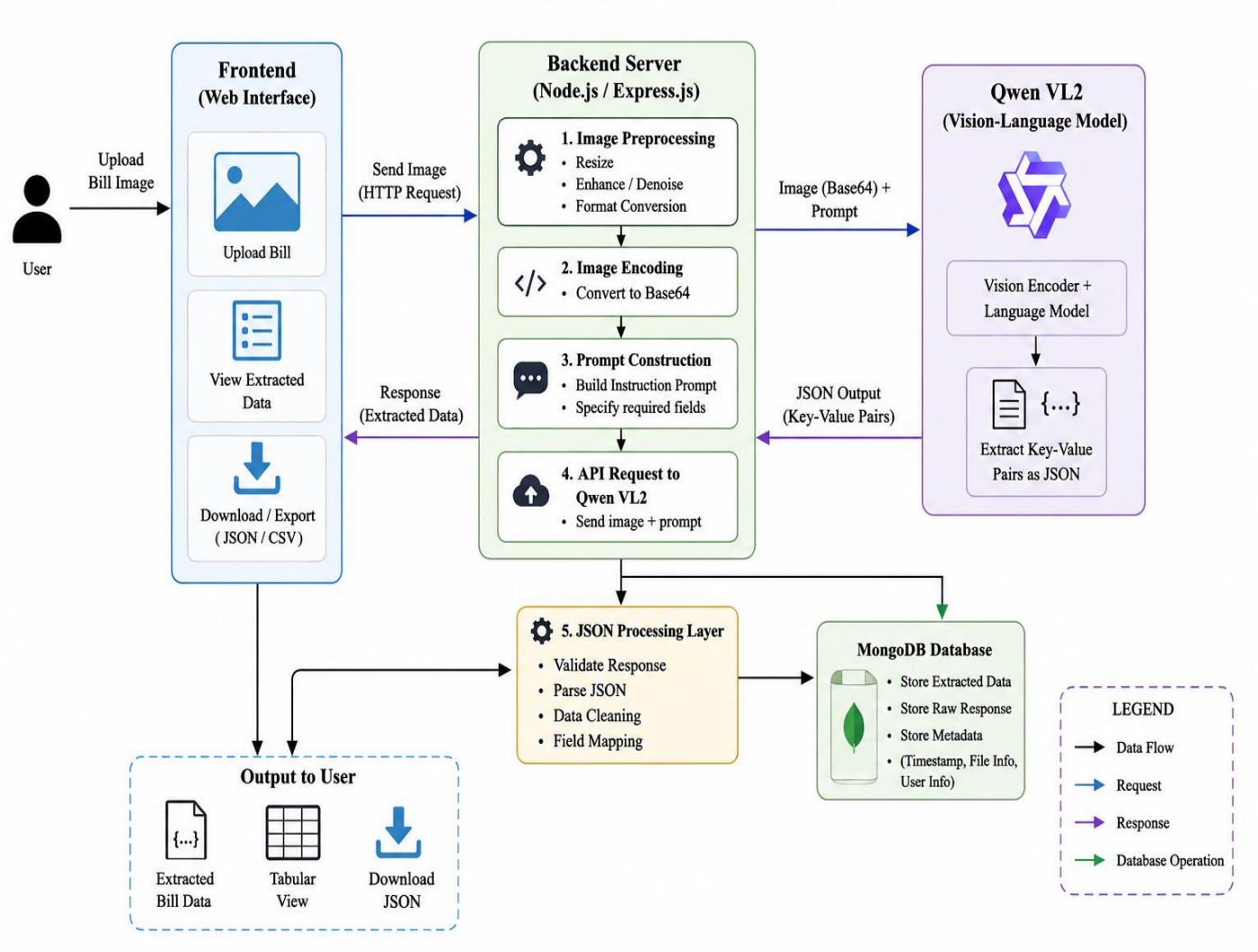


Fig. 1: System Architecture Diagram

**A. Frontend Interface:** A lightweight responsive web interface allows users to upload bill images in JPEG or PNG format. Client-side validation ensures acceptable file size and format before submission. The interface displays extracted fields immediately after processing.

**B. Node.js API Backend:** The backend is built using Express.js. On receiving an uploaded image, the server encodes it in base64 and constructs a structured prompt instructing Qwen VL2 to extract predefined fields. The model response is parsed, validated, and persisted to MongoDB. Error handling covers malformed JSON responses and low-confidence extractions.

**C. Qwen VL2 Inference:** Qwen VL2 receives the bill image and a system prompt specifying the output as a JSON object. The model processes the image holistically without a separate OCR stage, enabling robust extraction under moderate image distortions, varying fonts, and multi-language content. The model is accessed via an inference endpoint supporting base64-encoded image input.

**D. MongoDB Storage:** Extracted key-value pairs along with metadata (upload timestamp, filename, processing status, raw model output) are stored in MongoDB for downstream querying and analytics. Schema validation at the application layer enforces required fields and data types.

## V. DATA FLOW

The data flow of the proposed system begins when the user uploads a bill image through the frontend interface. The image is sent to the backend server, where preprocessing techniques such as resizing and enhancement are applied. The processed image is encoded and forwarded to the Qwen VL2 model along with a structured prompt.

The model extracts key-value information from the image and returns the output in JSON format. The backend processes and stores this data in the MongoDB database. Finally, the extracted information is sent back to the user and displayed in a structured format.

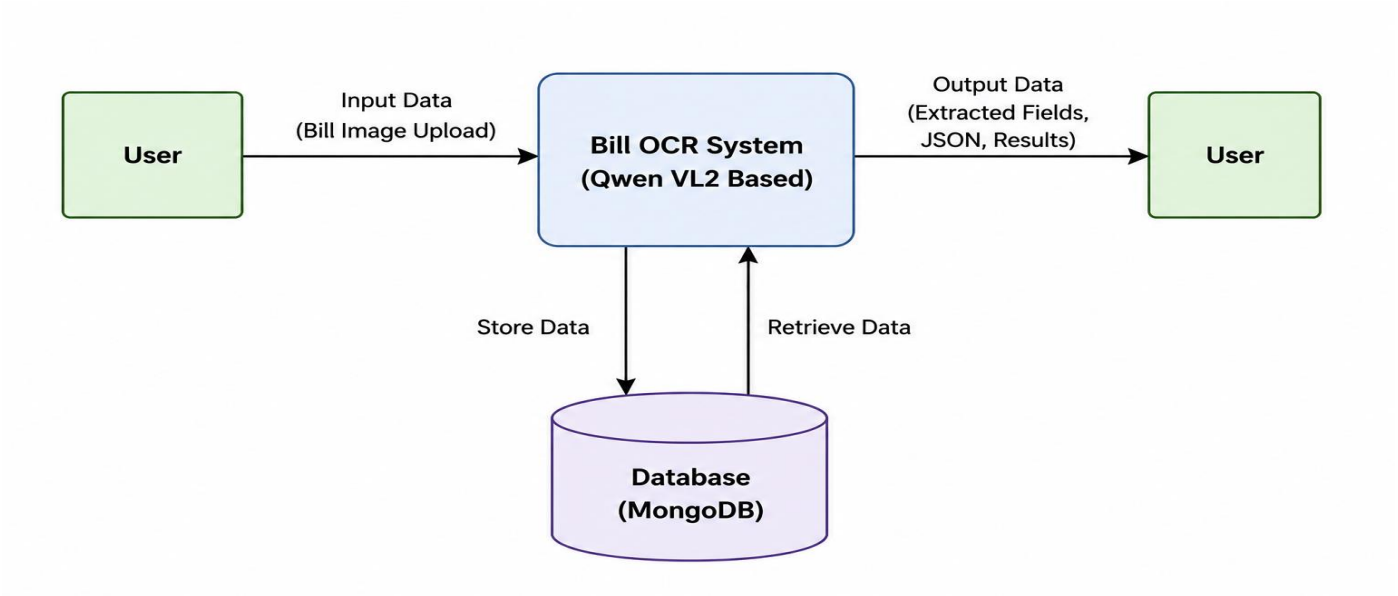


Fig. 2: Data Flow Diagram Level 0

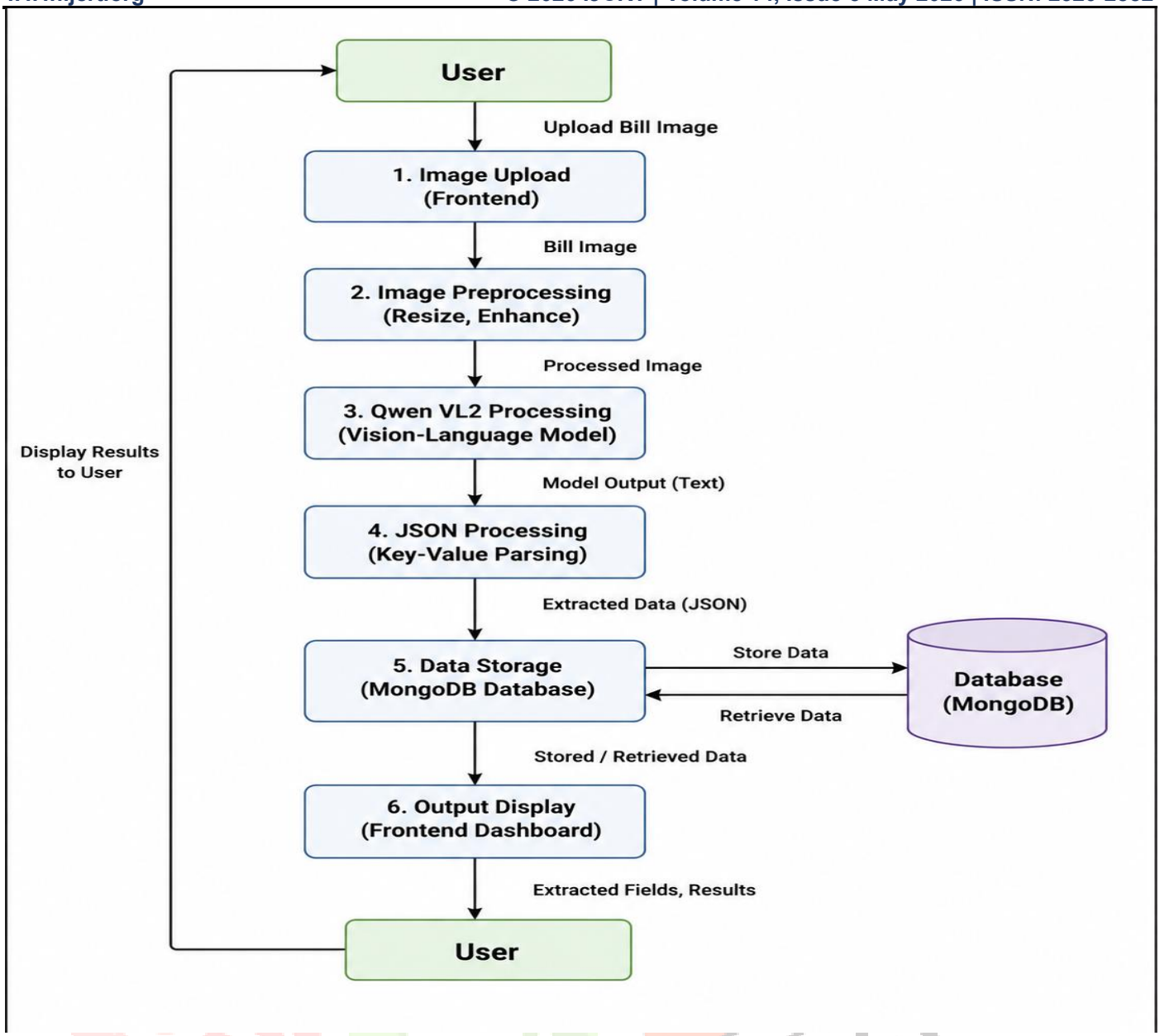


Fig. 3: Data Flow Diagram Level 1

## VI. METHODOLOGY

When a user uploads a bill image, the system performs lightweight preprocessing using the Sharp library: grayscale normalization and mild sharpening are applied to improve visual clarity without distorting layout information. The preprocessed image is encoded in base64 and embedded in a multimodal API request.

The system prompt engineering is central to performance. The prompt instructs: "Extract the following fields from this bill image and return a JSON object only: consumer_number, billing_period_start, billing_period_end, due_date, units_consumed, amount_due, taxes, total_payable. Set any missing field to null." This instruction-following behavior eliminates downstream regular-expression parsing entirely.

The JSON response from Qwen VL2 is lightly post-processed: numeric strings are cleaned of currency symbols and commas, and date strings are normalized to ISO 8601 format using a date parsing library. The structured record is stored in MongoDB alongside the raw model output for quality auditing and prompt iteration without re-inference.

For evaluation, each extracted field is compared against manually annotated ground truth. A field extraction is counted as correct if the normalized value matches the ground truth exactly (for numeric fields, within 1% tolerance for floating-point values). Overall accuracy is the macro-average across all fields and all bills.

## VII. EXPERIMENTAL RESULTS

The system was evaluated on 200 electricity bill images collected from five different utility providers across Maharashtra, India. Bills varied in layout, language (English and Marathi), resolution (72 DPI to 300 DPI), and scanning quality. All experiments were run on a server equipped with an NVIDIA T4 GPU.

**TABLE I**

**Field-Level Extraction Accuracy (%)**

| Field | Qwen VL2 (%) | Tesseract (%) |
|---|---|---|
| Consumer Number | 97.5 | 81.0 |
| Billing Period | 95.0 | 74.5 |
| Due Date | 96.0 | 78.0 |
| Units Consumed | 93.5 | 69.0 |
| Amount Due | 94.0 | 71.5 |
| Taxes | 91.0 | 65.0 |
| Total Payable | 95.5 | 72.0 |
| Overall Average | 94.6 | 73.0 |

Qwen VL2 consistently outperformed Tesseract-based extraction across all fields, achieving 94.6% overall accuracy versus 73.0% for Tesseract. The largest gains were in numeric fields (Units Consumed: +24.5%, Taxes: +26.0%) where Tesseract frequently confused digits due to font variations and print quality issues.

Average processing time for Qwen VL2 was 4.2 seconds per image on the GPU server, compared to 0.8 seconds for Tesseract. The accuracy advantage outweighs the latency cost for non-real-time batch processing. Error analysis showed that Qwen VL2 failures were predominantly on extremely low-resolution (under 96 DPI) or severely skewed images. Qwen VL2 also correctly handled bilingual (English-Marathi) bills where Tesseract produced garbled output.

## VIII. COMPARISON WITH RELATED WORK

Table II presents a feature comparison between the proposed system and existing OCR approaches for bill and invoice processing. The proposed system is unique in combining zero-shot extraction, multilingual support, and an open-weight deployable model.

**TABLE II**

**Comparison with Related Approaches**

| Feature | Tesseract+Regex | Cloud Vision API | LayoutLM | Proposed (Qwen VL2) |
|---|---|---|---|---|
| Zero-Shot | No | Partial | No | Yes |
| Fine-tuning Required | No | No | Yes | No |
| Multilingual | Partial | Yes | Partial | Yes |
| On-Premise Deploy | Yes | No | Yes | Yes |
| Overall Accuracy | 73.0% | ~91% | ~89% | 94.6% |
| Cost per Request | Free | Paid | Free | Free |

## IX. CONCLUSION AND FUTURE WORK

This paper presented an intelligent bill processing system using the Qwen VL2 vision-language model for structured key-value extraction from bill images. The system eliminates the fragility of traditional Tesseract-plus-regex pipelines by leveraging multimodal instruction-following capabilities. Experimental results demonstrate 94.6% overall field-level accuracy, outperforming the Tesseract baseline by over 21 percentage points across 200 real-world bills.

The Node.js and MongoDB backend provides a scalable, production-ready architecture suitable for enterprise deployment. The open-weight nature of Qwen VL2 enables on-premise deployment addressing data privacy requirements. Future work will explore: (1) fine-tuning Qwen VL2 on a domain-specific annotated dataset for further accuracy gains; (2) support for handwritten annotations on printed bills; (3) real-time streaming inference for sub-second latency; and (4) extension to other document types including medical reports and tax forms.

## ACKNOWLEDGMENT

The authors would like to thank Prof. Dr. A.K. Gupta, the Executive Director, DYPCET, and Prof. Dr. S.D. Chede, the Principal, DYPCET, for their forward-thinking guidance and support. The authors owe their gratitude to Dr. G.V. Patil, Head of the Department, CSE. The authors are extremely thankful to Prof. Swati Ambi, the project guide, for her continuous encouragement and technical direction throughout this research.

## REFERENCES

[1] R. G. Casey and E. Lecolinet, "A survey of methods and strategies in character segmentation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 18, no. 7, pp. 690-706, Jul. 1996.

[2] R. Smith, "An overview of the Tesseract OCR engine," in Proc. 9th IAPR Int. Workshop Document Anal. Syst., 2007, pp. 629-633.

[3] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 11, pp. 2298-2304, 2017.

[4] M. Li et al., "TrOCR: Transformer-based optical character recognition with pre-trained models," in Proc. AAAI, 2023, pp. 13094-13102.

[5] Y. Xu et al., "LayoutLM: Pre-training of text and layout for document image understanding," in Proc. 26th ACM SIGKDD, 2020, pp. 1192-1200.

[6] G. Kim et al., "OCR-free document understanding transformer," in Proc. ECCV, 2022, pp. 498-517.

[7] Qwen Team, "Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution," arXiv:2409.12191, 2024.

[8] A. Dengel and B. Klein, "smartFIX: A requirements-driven system for document analysis and understanding," in Proc. DAS, 2002, pp. 433-444.

[9] P. Harley et al., "Automated invoice key information extraction using deep learning," in Proc. ICDAR, 2021, pp. 1456-1461.