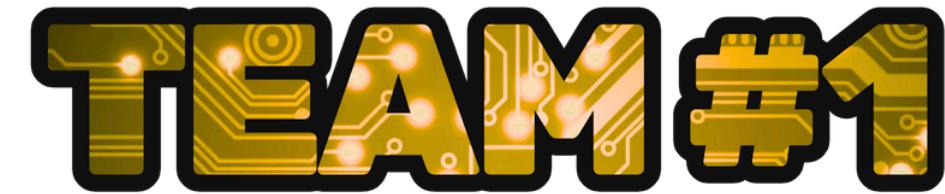


EMB GLOBAL

SMART DMS

AI-Powered Document Management System Architecture





THE COMPANY

Founded in 2018, EMB Global is an AI-powered B2B marketplace.

The company offers end-to-end digital transformation solutions such as:

- Cloud Services
- Software Development
- IT Consulting services
- Data driven Marketing Svcs

MARKET PRESENCE

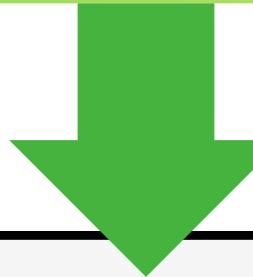
Privately held and based in Gurugram, India, EMB operates globally serving clients in over 23 countries with a focus on Southeast Asia and the Middle East. Their portfolio includes 30+ ready-to-deploy AI solutions.

KEY STRENGTHS

- AI + consulting for business problem-solving
- Smart execution and vendor management
- Scalable cloud and DevOps infrastructure
- Pre-built tools for rapid AI adoption

OPERATIONAL PAIN POINTS

EMB Global is a young startup



Process Inefficiencies

Immature internal processes cause recurring issues



Manual Processes

Reliance on manual tasks causes errors and delays



Communication Breakdown

Lack of communication between teams leads to delays



System Limitations

Current systems are inadequate for complex billing

Key Challenges

- **Frequent or duplicate contract requests** from finance staff
- **Complex RFPs** resulting from contract renewals
- **Delays in vendor payments** caused by manual notifications
- **Invoice discrepancies** arising from consumption-based billing
- **Uncertainty of purchase orders balance** (revenue forecast impact)
- **Client pushback** against separate service agreements
- **Expired Contracts and Service agreements** due to inaction

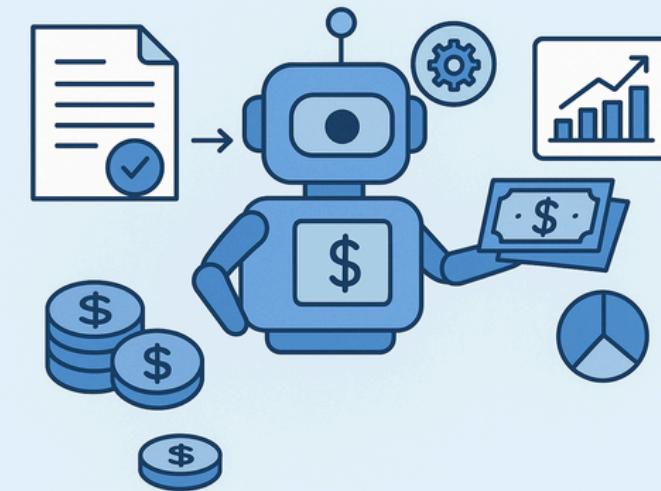
THE REQUEST FROM EMB

SMART DMS

An 'intelligent' Document Management System (DMS) fueled by AI to centralize the processing of Purchase Order, Contract, and Service Agreement documents. This system should also include a management dashboard.



AUTOMATED AP SYSTEM



An automation for their payment system to help administrate their Account Payables and Account receivables

DECISION

SMART DMS

An 'intelligent' Document Management System (DMS) fueled by AI to centralize the processing of Purchase Order, Contract, and Service Agreement documents . This system should also include a management dashboard.



Due to **time and resource** constraints, and the **scope** of this course, we've agreed to focus only on their top priority request and enrich it with LLM powered features (chatbot)

SOLUTION PART 1: THE AI-POWERED DMS

The AI-Powered DMS works as follow:

Ingest & Read

- Employees upload any document (PO, invoice, contract).
- Our system uses OCR to instantly read and extract key data (e.g., PO Number, Expiration Date, Amount) and LLM model for chatbot

Validate & Store

- The system automatically cross-checks data (e.g., Does this invoice match its PO?).
- All validated data is stored in one secure, central database.

Alert & Monitor

The system sends proactive alerts to the right teams before problems happen. Such as:

-  Alert: "PO #12345 is 80% depleted."
-  Alert: "Contract with ABC Corp expires in 60 days."

The DMS will help prevent costly errors, like payments on expired contracts, and save hundreds of working hours for finance and operations teams.

SOLUTION PART 2: AI CHATBOT

The friendly chatbot will assist employees in various areas, including:

Checking PO Status

- Employee: "What's the balance on PO-123?"
- Chatbot: "The remaining balance on PO-123 for Vendor X is \$10,000."

Checking Contract Deadlines

- Employee: "When does our contract with ABC Corp expire?"
- Chatbot: "The service agreement with ABC Corp expires on June 30, 2026."

Getting AI-Powered Summaries

- Employee: "Summarize the ABC Corp agreement."
- Chatbot: "It's a 2-year, \$100k contract for cloud services that auto-renews. The cancellation notice is 90 days."

These sample conversations with the chatbot show how it empowers employees with self-service information, reducing interruptions and allowing teams to focus on strategic work.

STRATEGIC AND BUSINESS BENEFITS

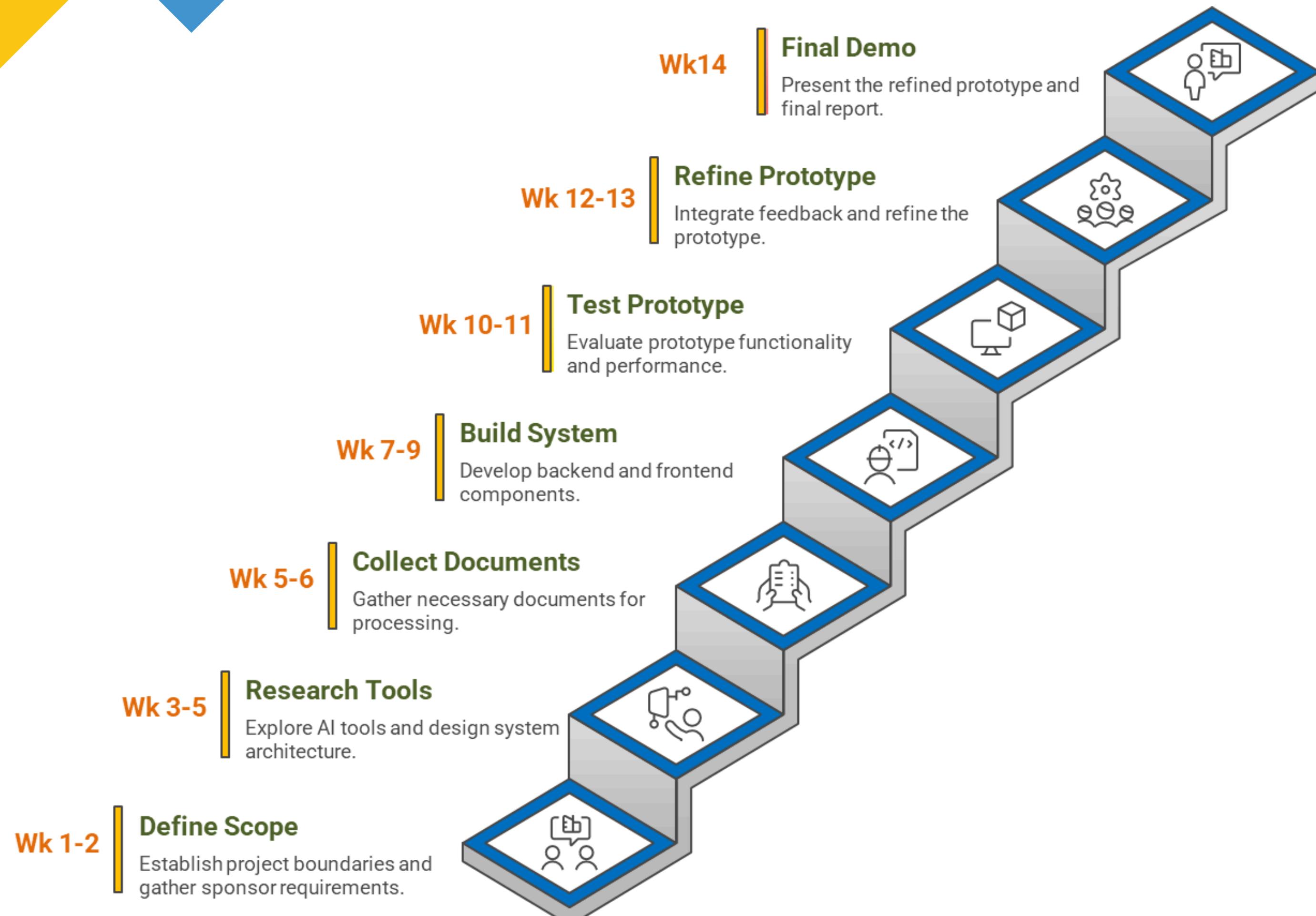
Strategic Impact:

Depending on current and upcoming challenges, this system will empower EMB Global to grow and thrive by replacing manual business processes with automation. It ensures scalability, efficiency, and accuracy, enabling the company to handle increasing workloads without additional staffing.

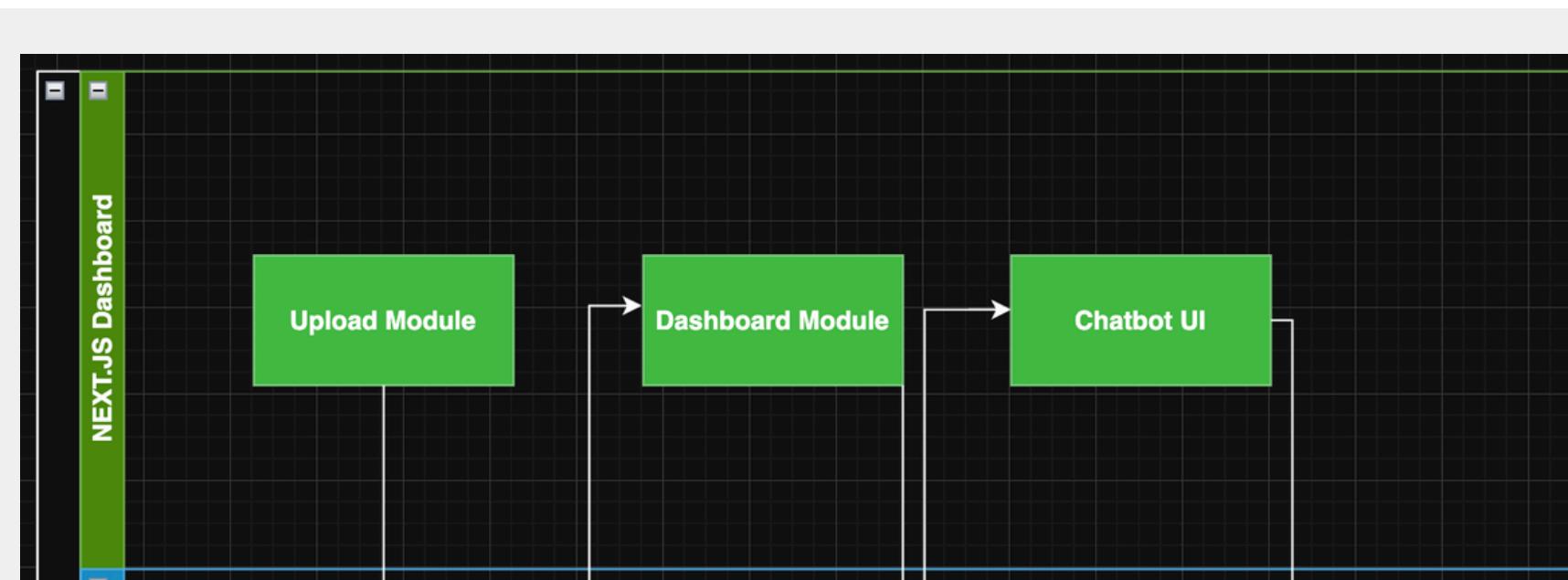
Key Benefits:

- Increase the productivity
- Cost Efficient
- Reduce Costly Errors
- Boost Company Growth.

PROJECT TIMELINE

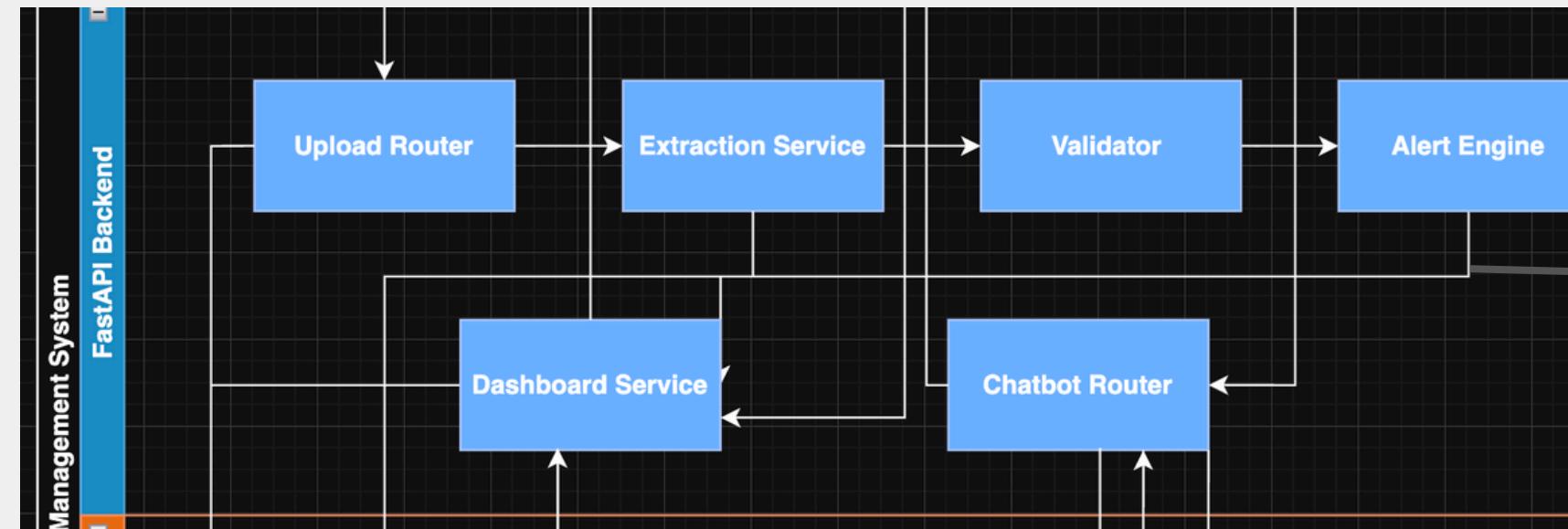


SOLUTION ARCHITECTURE



LEVEL 1

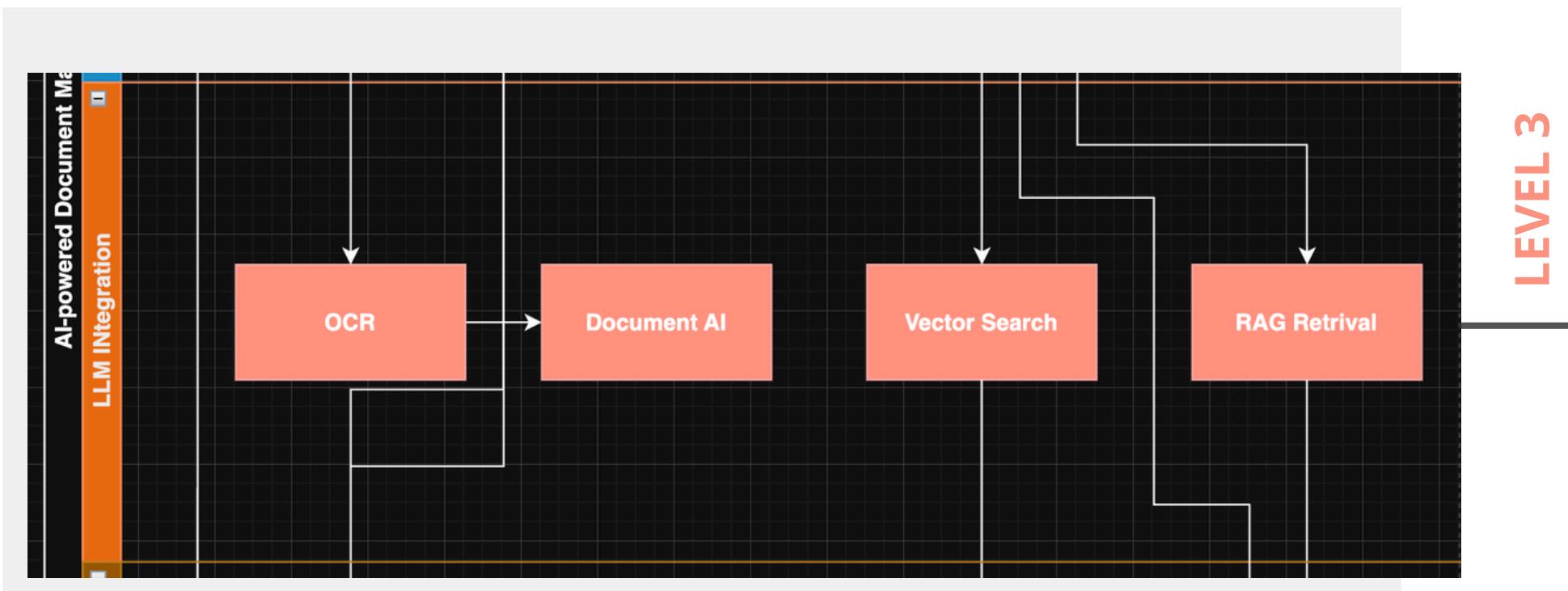
- User-facing web application with three modules
- Upload Module: Multi-file PDF upload interface
- Dashboard Module: Real-time metrics and alerts display
- Chatbot UI: Conversational document query interface



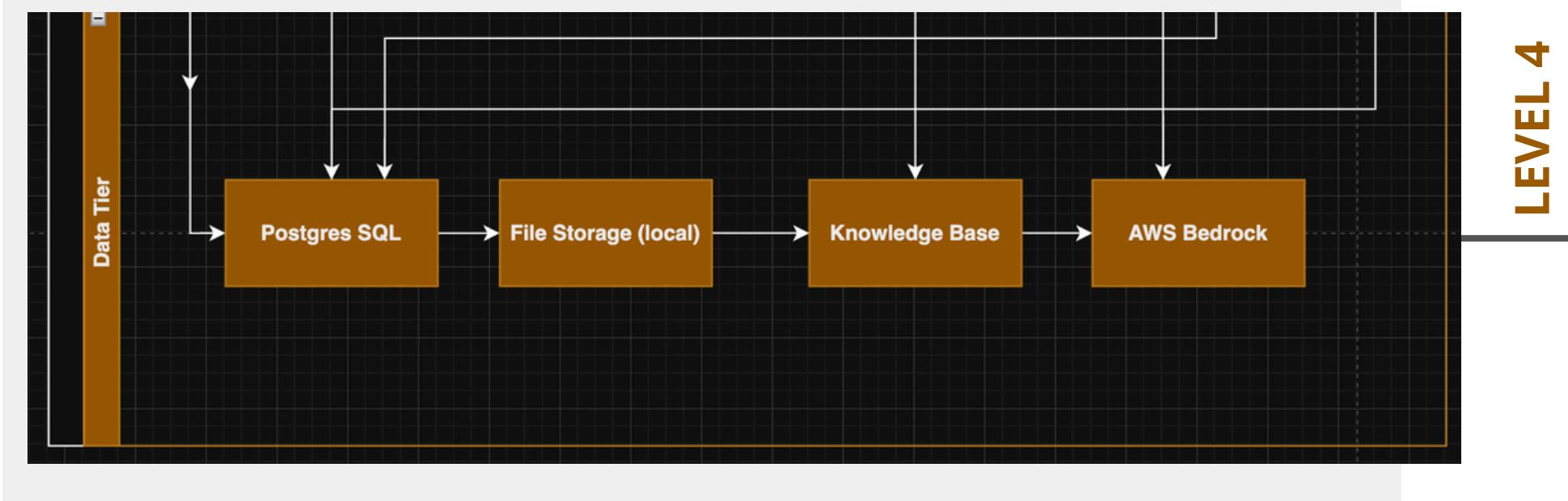
LEVEL 2

- RESTful API handling all business logic
- Upload Router → Extraction Service → Validator → Alert Engine
- Dashboard Service: Aggregates data for frontend
- Chatbot Router: Processes natural language queries

SOLUTION ARCHITECTURE



- **OCR:** Text extraction from PDFs
- **Document AI:** Structured field extraction (Claude through AWS Bedrock)
- **Vector Search:** Semantic similarity matching
- **RAG Retrieval:** Context-aware response generation



- **PostgreSQL:** Structured document & alert storage
- **File Storage:** Raw PDF repository (local/GCS)
- **Vector Database:** Document embeddings for chatbot
- **AWS Bedrock:** LLM for extraction & conversation

TECH STACK

Frontend Layer

- **Framework:** Next.js / React
- **UI Components:** Custom dashboard components
- **File Upload:** Multi-file PDF upload interface
- **Real-time Updates:** WebSocket/polling for alert notifications

Backend & API

- **API Framework:** FastAPI (Python 3.10+)
- **Architecture:** RESTful microservices
- **Authentication:** JWT-based auth
- **Cloud Storage :** S3 bucket

AI/ML Services

- **Document OCR:** Amazon Textract
- **LLM:** Claude 3 Haiku (AWS Bedrock)
- **Document Understanding:** AWS Bedrock Document AI
- **Vector Embeddings:** Amazon Bedrock Embeddings
- **RAG Framework:** LangChain / Custom RAG pipeline

TECH STACK

Data Storage

- **Relational Database:** PostgreSQL
- **Vector Database:** Amazon OpenSearch / Pinecone
- **File Storage:** AWS S3
- **Cache Layer:** Redis (optional)

Cloud Infrastructure

- **Compute:** AWS Lambda (serverless functions)
- **Monitoring:** AWS CloudWatch
- **API Gateway:** AWS API Gateway
- **Security:** AWS IAM, KMS for encryption

DevOps & Tools

- **Version Control:** Git / GitHub
- **CI/CD:** GitHub Actions / AWS CodePipeline
- **Containerization:** Docker
- **Development:** Cursor (AI-assisted IDE)
- **Testing:** Pytest, Postman
- **Monitoring:** CloudWatch

DOCUMENT UNDERSTANDING: TEXTTRACT

TEXTTRACT reads.



Step 1

AWS Textract

Extracts text & key-value pairs
from PDFs

Input: Invoice.pdf

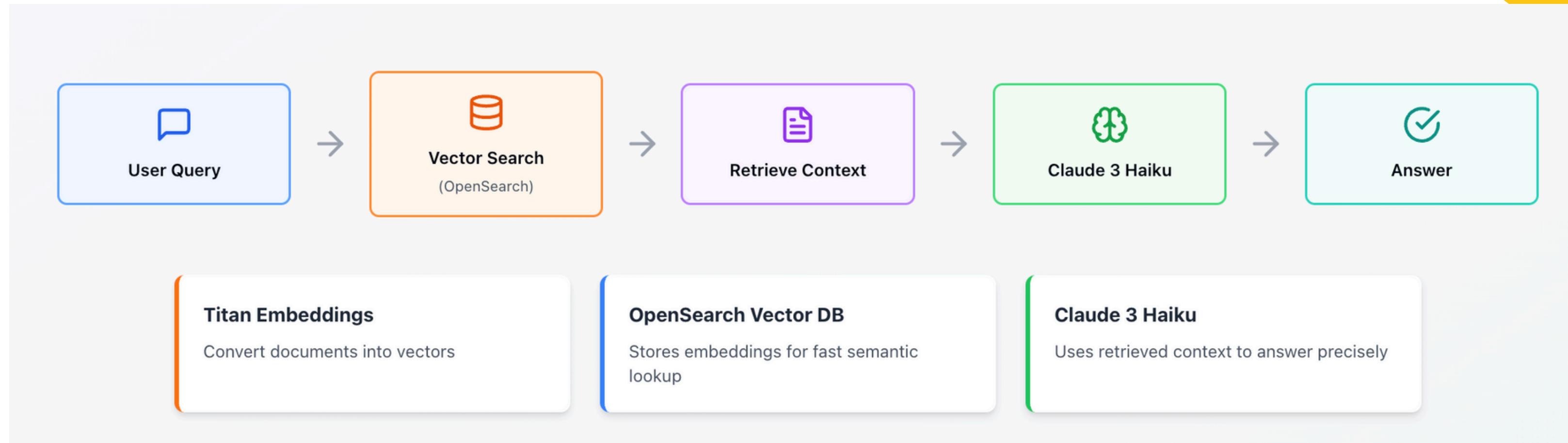
OCR Output →

"Vendor: MANTARAV DIGITAL ... Total AED 6,300"

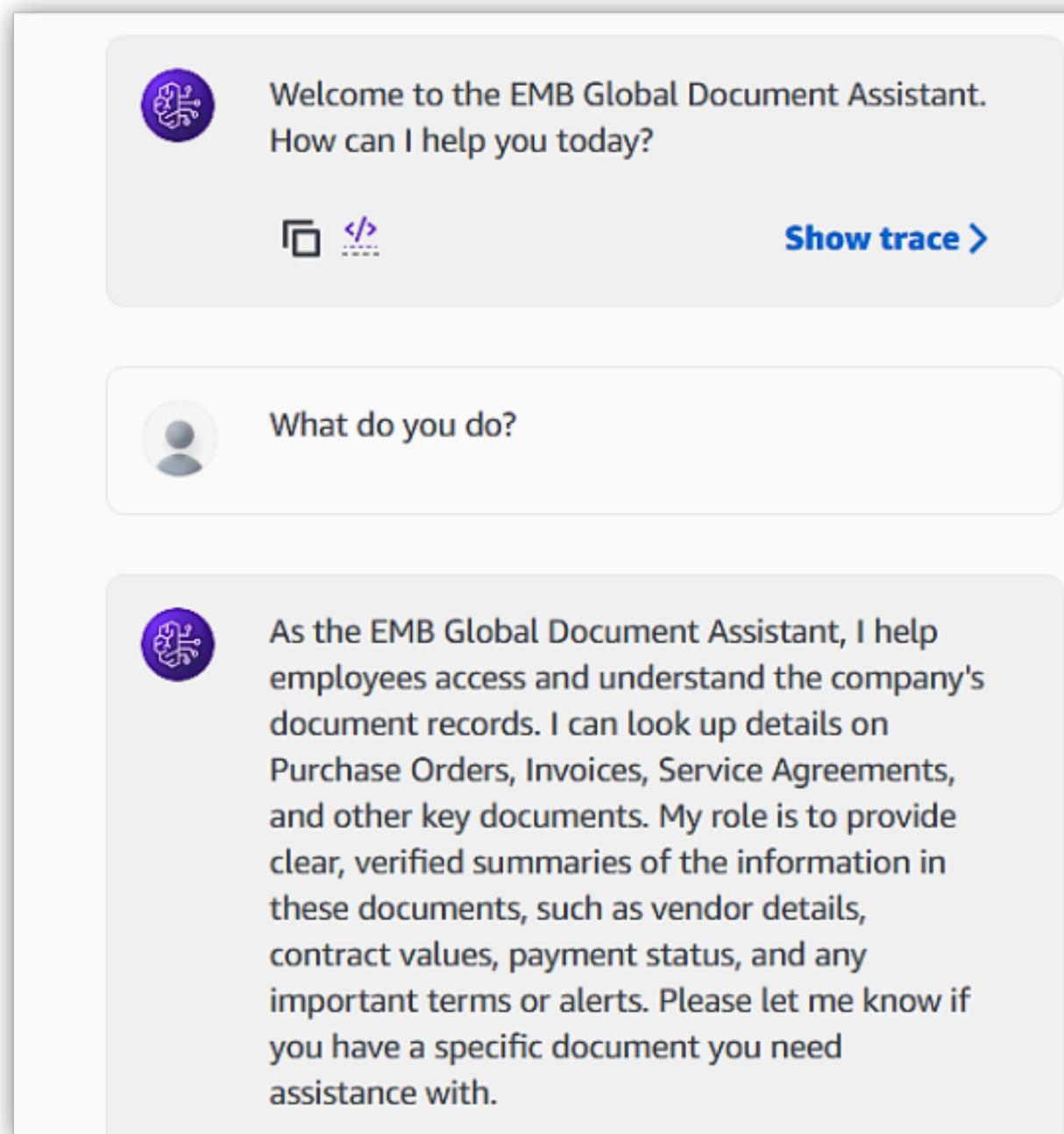
Claude 3 Haiku (AWS Bedrock) →

```
{  
  "Vendor": "Mantarav Digital",  
  "P0_Number": "1282689",  
  "Total": 6300,  
  "Status": "Valid - within P0 cap"  
}
```

RETRIEVAL-AUGMENTED GENERATION (RAG) PIPELINE

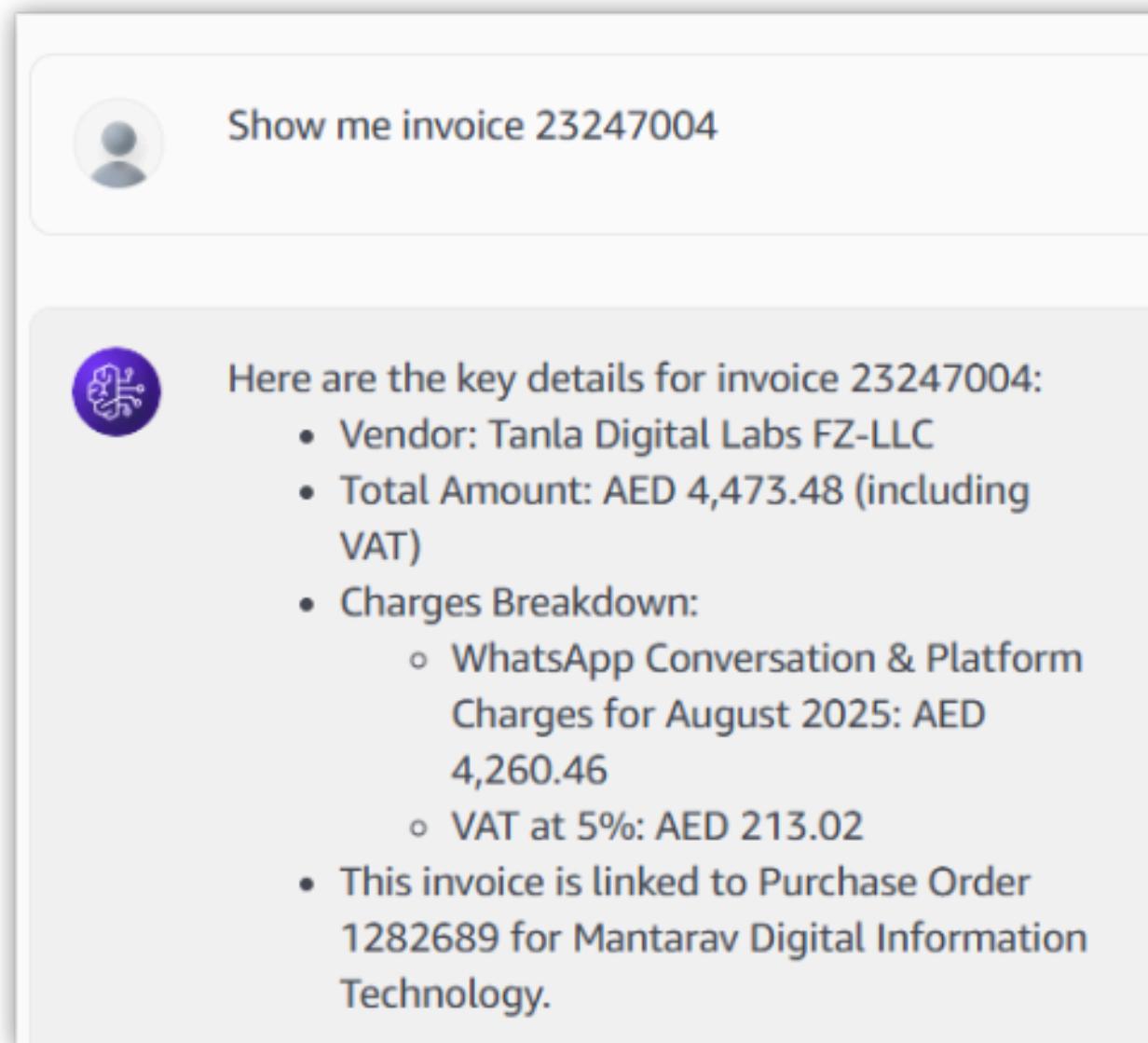


CHATBOT USING AWS BEDROCK: AGENT OVERVIEW



- Developed using Amazon Bedrock Agents with Claude 3 Haiku
- Integrates company knowledge base for invoice and PO queries
- Custom system instructions ensure consistent professional tone
- Responds contextually with verified document information

CHATBOT USING AWS BEDROCK: INVOICE PROCESSING



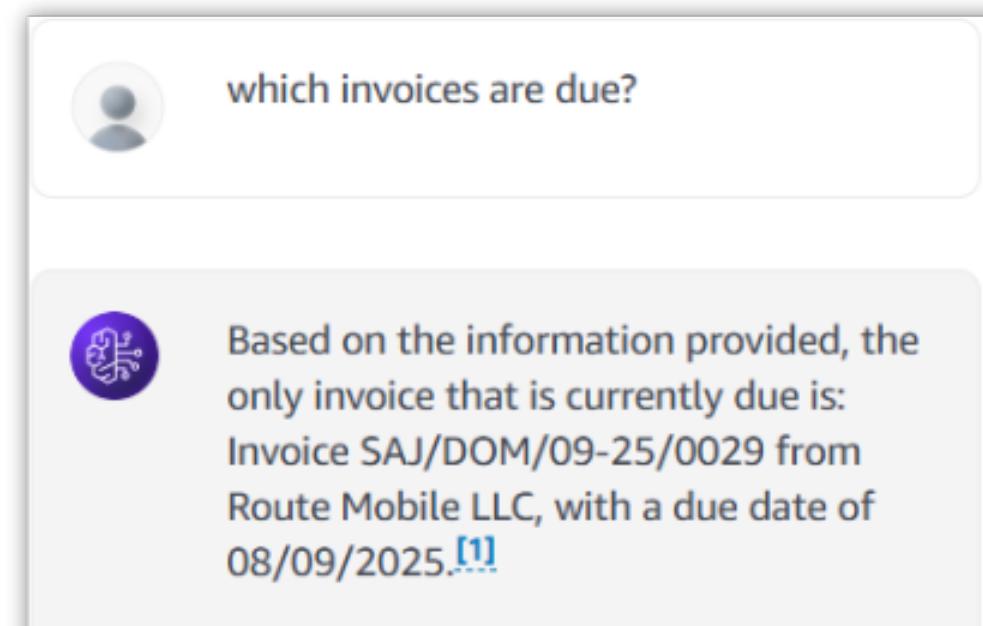
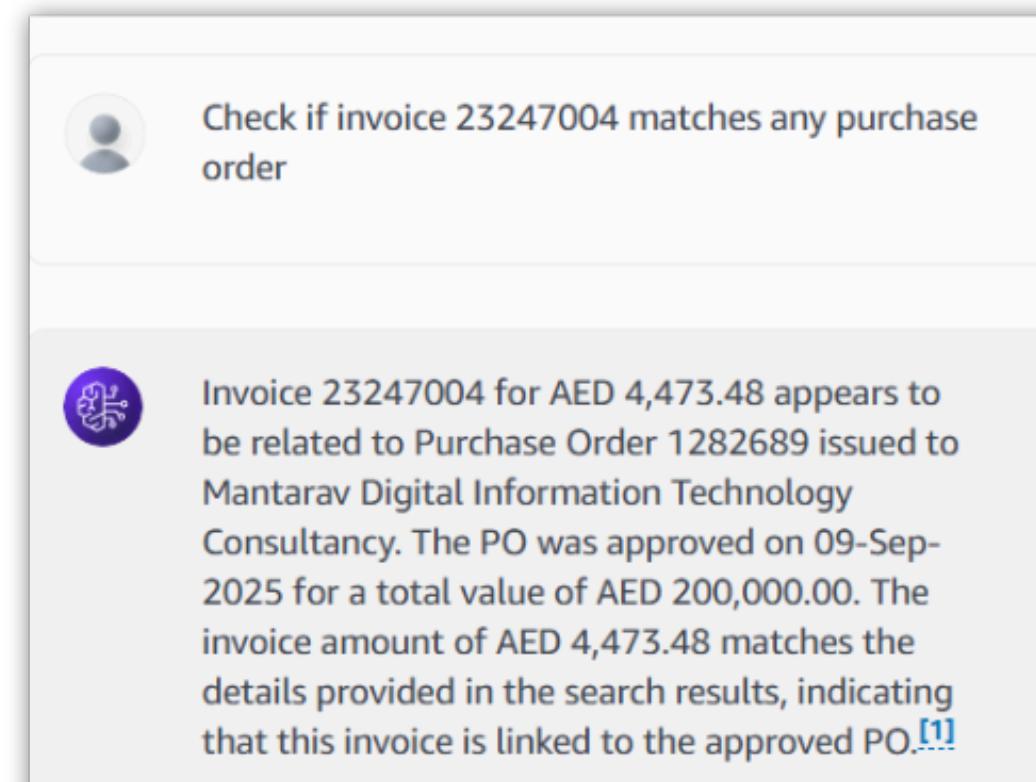
A screenshot of a chatbot interface. On the left, there's a user input field with a person icon containing the text "Show me invoice 23247004". On the right, the bot responds with a purple icon containing a gear symbol, followed by the message: "Here are the key details for invoice 23247004:" and a bulleted list of invoice details.

Here are the key details for invoice 23247004:

- Vendor: Tanla Digital Labs FZ-LLC
- Total Amount: AED 4,473.48 (including VAT)
- Charges Breakdown:
 - WhatsApp Conversation & Platform Charges for August 2025: AED 4,260.46
 - VAT at 5%: AED 213.02
- This invoice is linked to Purchase Order 1282689 for Mantarav Digital Information Technology.

- Extracts structured details from uploaded invoices
- Retrieves linked Purchase Orders for cross-verification
- Detects mismatches between invoice and PO totals
- Provides quick insights on vendor, date, and payment status

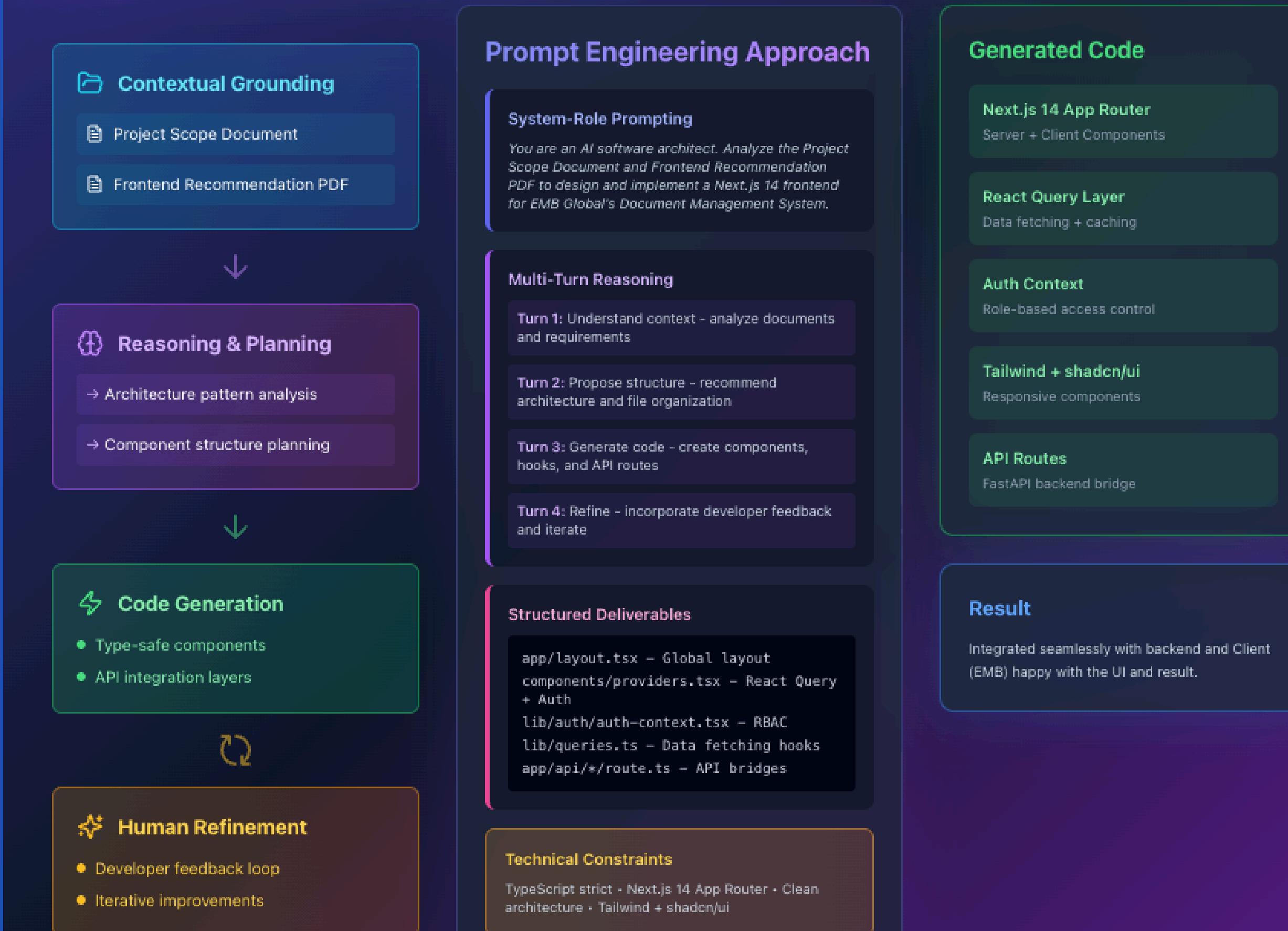
CHATBOT USING AWS BEDROCK: PO & CONTRACT INSIGHTS



- Summarizes Purchase Order balances and linked invoices
- Identifies expiring Service Agreements and flags alerts
- Operates via integrated Knowledge Base (RAG) workflow
- Enables finance teams to query and validate records instantly

AI-Assisted Frontend Generation Workflow (Codex through Cursor)

Contextual Grounding → Reasoning & Planning → Code Generation → Refinement



FRONTEND: USING CURSOR/CODEX

The screenshot displays a web application interface for a Document Management System (DMS) dashboard. The top navigation bar includes a back button, forward button, address bar (localhost:8000), and a message about a new Chrome available update. The user profile at the top right shows "Isha Pathak FINANCE".

Operations Pulse
Track intake velocity, spending utilization, and outstanding exceptions in real time.

Quick Document Upload
Drop PDF documents here to start processing POs, invoices, and agreements.

Drop PDF documents here
Supports POs, invoices, agreements • Max 25MB per file

DMS Assistant
Ask about spending, alerts, or document status.

Hi! I'm the DMS assistant. Ask about purchase orders, invoices, agreements, or alerts and I'll guide you.

What's the status of PO-2024-001? You

Purchase orders are tracked with cap, utilization, vendor, and expiry metadata. You can review PO balances on the dashboard or open the document details for more information.

What's the status of PO-2024-001?

How do alerts get triggered?

Summarize the Field Support agreement.

Ask something about the DMS... Send

Quick Actions
Navigate to key sections of your DMS dashboard.

Documents → View all documents

Exceptions → Review issues

Alerts → View notifications



Document Inventory

Central repository of purchase orders, invoices, and agreements.

Search by ID, client, ver

All categories



Drop PDF documents here

Supports POs, invoices, agreements • Max 25MB per file

ID	Title	Category	Client/Vendor	Amount	Status	Confidence	Linked
PO-2024-001	EMB Retail Supply PO Created 3/2/2024	CLIENT PO	EMB Retail	USD 150,000	APPROVED	<div style="width: 98%;">98%</div>	AGR-2024-002
INV-2024-032	Supplier Invoice - Batch #32 Created 3/11/2024	VENDOR INVOICE	EMB Retail Vendor: Northwind Components	USD 28,500	PENDING REVIEW	<div style="width: 91%;">91%</div>	PO-2024-001
AGR-2024-002	Service Agreement - Field Support Created 1/17/2024	SERVICE AGREEMENT	EMB Logistics Vendor: Helios Services	USD 56,000	APPROVED	<div style="width: 95%;">95%</div>	Unlinked
INV-2024-045	Client Invoice - Retail Expansion Created 3/15/2024	CLIENT INVOICE	EMB Retail	USD 72,000	FLAGGED	<div style="width: 76%; background-color: #f08080;">76%</div>	PO-2023-014
PO-2023-014	Vendor Supply PO Created 11/21/2023	VENDOR PO	EMB Retail Vendor: Helios Services	USD 90,000	APPROVED	<div style="width: 99%;">99%</div>	Unlinked

Core Generative AI Concepts Behind Our Workflow

- Instruction-Tuned LLM
(Codex)

- Prompt Engineering

- System-Role Prompting

- Contextual Grounding

- Multi-Turn Reasoning
(Chain-of-Thought?)

- Human-in-the-Loop
Refinement

* Result *

- ✓ Frontend integrated seamlessly with the initial backend

- ✓ UI matched design specifications

- ✓ EMB Global team happy with UI and design

- ✓ Demonstrated the feasibility of AI-assisted frontend generation in a real business context

Risk Assessment & Mitigation Strategy

Risk Category	Description	Impact	Likelihood	Mitigation Strategy
Data Extraction Inaccuracy	Textract/LLM may miss or misread key invoice/PO fields	High	Low	Textract for OCR + Claude for reasoning & field normalization
Mismatch Alert Errors	Incorrect PO-Invoice alerts due to imperfect field interpretation	High	Low	Improve extraction recall (flag uncertainty), validate on sample cases
Document Format Variability	Vendor layouts differ widely; new templates cause extraction errors	Medium	High	Template-agnostic extraction handled by AWS Textract + Claude
Security & Access Control	Sensitive docs accessed by wrong users	High	Low	Role-based access control
User Adoption Risk	Employees may not fully adopt the workflow	Medium	Low	Intuitive UI + quick onboarding
Scalability Constraints	High volume may slow system	Medium	Low	Low likelihood due to AWS-native pipeline; scaling handled by client dev team

AI PERFORMANCE EVALUATION METRICS

Document Extraction Metrics

Field Recognition Accuracy **93–96%**

Data Validation Accuracy **88–94%**

Retrieval (RAG) Metrics

Retrieval Precision **90–95%**

Answer Relevance **90–95%**

Chatbot Performance Metrics

Query Success Rate **90–96%**

Response Time **0,7–1,2s**

User Satisfaction Score **95%+**

LLM Summary & Quality

Summary Accuracy **88–94%**

Hallucination Rate **3–5%**

BUSINESS CASE

Current Situation (Baseline)

- Supplier network: 1,000+ partners (\approx 50 active at any given time)
- Customer base: \approx 200 clients with multiple concurrent service agreements
- Transaction load: \sim **1000 documents** per month (PO, invoices, contracts, service agreements)
- Manual effort: 10 FTEs managing POs, invoices, and contracts across fragmented tools

Solution Costs

Cost Component	Assumption	Annual Cost
AWS Infra (Yearly)	\$260/month (details in Annex 1)	\$3,120
Development and deployment	\sim 320hrs @ \$70/hr (internal + consulting)	\$22,400
Maintenance and support	10–12 hrs/month @ \$50/hr	\$7,200
Total Estimated Cost		\$32,720

Solution Quantitative Impact

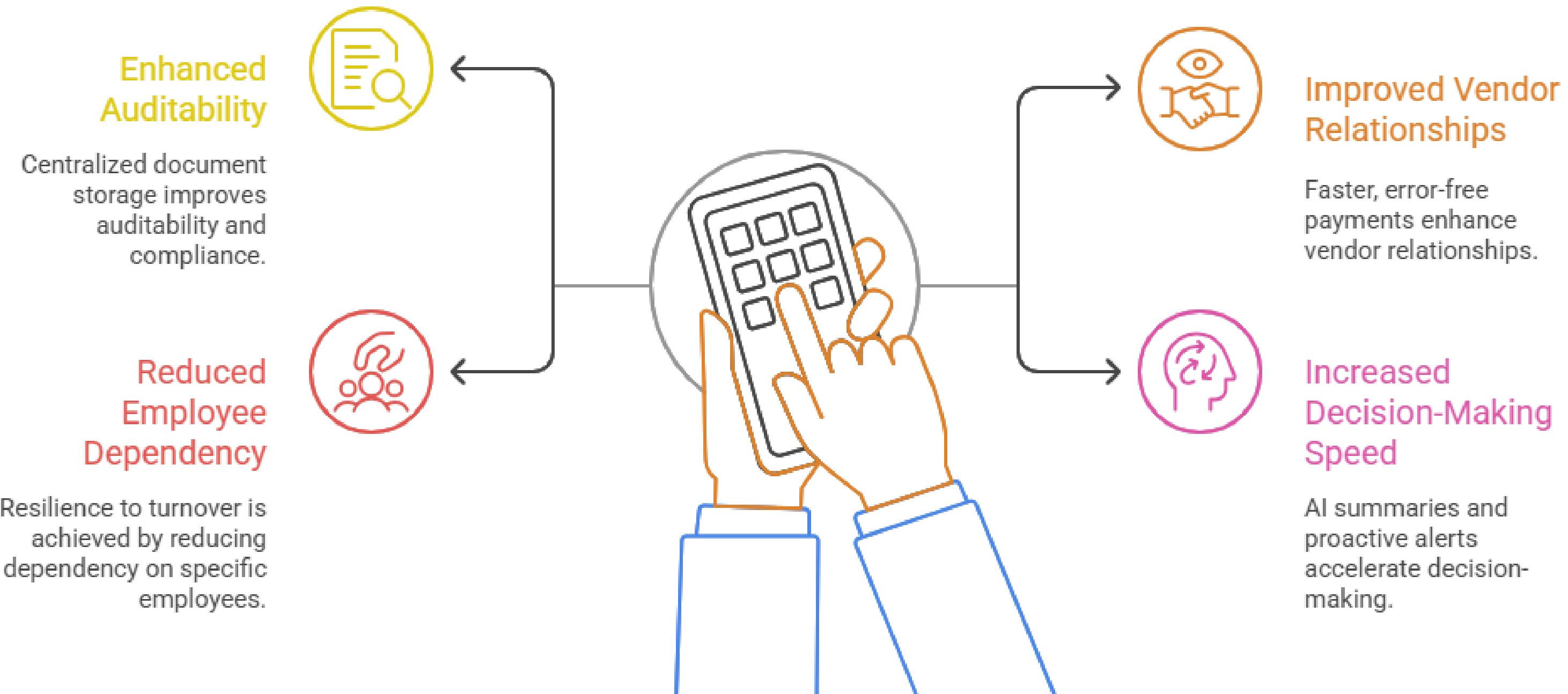
Benefit Area	Estimated Annual Savings
3 FTE workload eliminated (30% reduction)	\$72,000
Faster invoice cycle (late fees avoided) 5 days \rightarrow 1 day	\$ 5,000
Error reduction in invoices/contracts (10% to 2%)	\$8,000
Improved compliance (contract renewal leakage)	\$10,000
Total Annual Savings (Year 1)	\$ 95,000

ROI

Total Savings \sim **\$95k**
Total Ops Costs \sim **\$33k**
Net Annual Benefit: **\$ 62k**
ROI: **188%**
Payback Period: **4.2 months**

BUSINESS CASE

Solution Qualitative (Strategic) impact



DEMO SCENARIO

Step 1: Uploading Documents

An employee begins by uploading a new document to the DMS. The document can be of any type, such as a Client Purchase Order (PO), invoice, Master Service Agreement (MSA), etc...

Step 2: Document Processing

Once the document is uploaded:

- The system reads and processes the document to extract relevant information such as the PO number, MSA number, issue date, expiration date, and any other relevant metadata.
- The document is then securely stored in the system, categorized, and indexed for easy retrieval at a later time.
- If the document is identified as a PO or Invoice with no link to another (either through the PO number or MSA number), an alert is triggered. This alert is displayed on the dashboard and sent through the system to notify the relevant employee.

DEMO SCENARIO

Step 3: Using the Chatbot for Information Retrieval

- For instance, the employee can ask the chatbot for a summary of the document, its issue date, or expiration date.
- The Chatbot(Powered by Claude Haiku) responds accurately with real-time information extracted from the document, saving the employee time from having to manually search through documents.

Step 4: System Notifications and Alerts

The system monitors document activity and validation thresholds:

- The system helps to automate most of the processes and surfaces real-time notifications if any issues arise.

DEMO SCENARIO

Step 5: Dashboard Overview

The system provides a comprehensive dashboard to the employee, with several key metrics and features displayed:

- Active Client POs: Displays the number of active Client POs, showing if there has been an increase or decrease in comparison to the last 30 days.
- Invoice Utilization: Shows the percentage of PO caps consumed, indicating how much of the available PO limit has been utilized.
- Exceptions: Displays any open validation issues or exceptions that may need human-in-loop review.
- Avg. Processing Time: Shows the average time taken for the system to process and validate each document from ingestion to validation.
- Monthly Document Activity: A graph showcasing the number of documents processed each month, segmented by Client and Vendor documents.
- Documents by Type: A visual pie chart showing the types of documents ingested, with categories like Client POs and Vendor Invoices

DEMO SCENARIO

Step 5: Viewing Recent Documents

The system also provides quick access to recently processed documents, allowing employees to view the latest additions, such as contracts, POs, or vendor invoices, and understand their statuses.

The screenshot displays the EMB GLOBAL DMS Console interface. On the left, a sidebar contains links for Dashboard, Documents (Active), Exceptions, Alerts, and Settings. The main area is titled "Document Inventory" and describes it as a "Central repository of processed documents with OCR content". It shows a single document entry:

Category	Details
Date	date unless otherwise specified
Type	Client PO
From	AI Futtaim Private Company LLC
On	2025-12-03
To	MANTARAV DIGITAL INFORMATION TECHNOLOGY
Notes	MASTER SERVICE AGREEMENT MSA Number: MSA-2025-002 AGREEMENT PARTIES SERVICE PROVIDER: MANTARAV DIGITAL INFORMATION TECHNOLOGY CONSULTANCY SOLE PROPRIETORSHIP L. L. C

At the bottom left, a callout box says "Need faster ingest? Sync WhatsApp, Gmail, and Drive ingest endpoints in Settings to fully automate intake." The top right corner shows the user profile "Isha Pathak FINANCE".

LESSONS LEARNED: FRONTEND AI-ASSISTED DEVELOPMENT

- Structured prompting beats random prompting

Using contextual grounding (scope document + frontend recommendation PDF) + role-based prompting produced far better outputs than generic “build me a dashboard” prompts.

- Codex-style LLMs excel with code-heavy tasks

Fine-tuned code models generated consistent Next.js + Tailwind + shadcn components, API routes, and React Query layers with minimal correction compared to general-purpose LLMs.

- A defined workflow matters more than model size

The 4-stage cycle – Contextual Grounding → Reasoning/Planning → Code Generation → Human Refinement created reliable, production-ready frontend code.

- Structured deliverables accelerate iteration

Asking the model to output file-by-file structures (/app, /components, /lib) eliminated ambiguity and made the generated code directly usable.

- Human refinement remains essential

Developer review + iterative improvements were key for polishing UI logic, hooks, and state handling, but the base scaffolding was generated extremely efficiently.

LESSONS LEARNED: DOCUMENT PROCESSING AND AI CHATBOT

- Some level of standardization of documents needed from client side for the app to work prop lots of documents don't have the company name, lots of Invoices don't have PO numbers, EMB need to standardize docs

- Smooth integration with the AI-generated frontend

Because the frontend followed consistent AI-generated API patterns (React Query + clean REST routes), connecting FastAPI endpoints to UI components was straightforward and reliable.

- Huge variation in document formats

Even within the same category (POs, invoices, RFPs), layouts and structures were inconsistent, making extraction unpredictable.

- Scanned / image-based PDFs caused some OCR inconsistencies

Low-quality scans, rotated pages, and embedded images caused missing or distorted fields.

- Claude Haiku's lower reasoning capability required strict instruction lists

Haiku was chosen for low cost and fast inference, but it needed very explicit extraction prompts and rigid JSON schemas to avoid hallucinations or missing fields.

- Textract Significantly Outperformed Tesseract in Field Extraction

In our extraction tests, AWS Textract achieved 90%+ accuracy, compared to ~60-70% with Google Tesseract, especially for tables, key-value pairs, and irregular invoice/PO layouts.

- Chatbot performance still answered 90%+ questions accurately

Even with imperfect extraction, the chatbot answered document queries and generated structured summaries with near-perfect accuracy, thanks to strong retrieval and clean metadata.

FUTURE WORK

- Enhance the extraction pipeline with stronger models

Texttract is already implemented, but reliability can be improved by pairing it with higher-accuracy LLMs (e.g., Sonnet/Opus) for reasoning, field correction, and validation, reducing inconsistencies that Haiku sometimes misses.

- Hybrid model routing (Haiku → Sonnet)

Keep Haiku for fast/low-cost passes, but route complex or ambiguous documents to Sonnet for improved accuracy, reasoning, and mismatch detection.

- Restructure document storage & metadata schemas

Introduce richer metadata fields, standardized identifiers, and improved embedding organization to boost chatbot retrieval accuracy and system scalability.

- Direct WhatsApp integration for document ingestion

Allow invoices and POs shared over WhatsApp to be ingested automatically into the DMS, triggering OCR, extraction, and alert workflows.

- Client-driven document standardization

EMB Global are working on making document formats more consistent and ensuring all files include essential details (PO number, invoice number, vendor info), which will significantly enhance model performance.

CODE:

https://github.com/rishabh15b/DMSDashboard_backend.git

https://github.com/rishabh15b/DMSDashboard_frontend.git

ANNEX 1

AWS Infrastructure Cost Calculation (Year 1)

Cost Category	Assumption	Monthly	Annual Cost (\$)
Document extraction use	Avg 2,000 tokens per document round-trip (input + output) for 1,000 documents	\$ 4	\$ 48
Chatbot use	200 queries/day = 6,000/month. Average 1,000 tokens/interaction	\$ 12	\$ 144
Bedrock LLM		\$ 16	\$ 192
OCR – Textract	Textract: ~\$1.50 per 1,000 pages. Assume 2 pages/document → 2,000 pages/month	\$ 3	\$ 36
Vector Database	OpenSearch Serverless (~\$120–\$160/month for low tier)	\$ 140	\$ 1,680
S3 storage + GET/PUT	1,000 PDFs/month @ 300 KB → ~300 MB added. Yearly < 5 GB	\$ 5	\$ 60
PostgreSQL (RDS)	Small production-grade instance: db.t3.micro + storage	\$ 30	\$ 360
Lambda + API Gateway	Low volume usage \$15/month	\$ 15	\$ 180
Frontend Hosting	Amplify or S3	\$ 5	\$ 60
CloudWatch Monitoring		\$ 8	\$ 96
Risk Contingency	10% of total cost	\$ 22	\$ 264
Total Annual AWS Infra cost estimation			\$ 3,120



THANK YOU