JOSIAH STEVE SAM

AI Engineer | Agents, RAG & Document AI | Chennai, India (Open to Relocation)
+91 7373077757 | josiahsteve29@gmail.com | Portfolio | LinkedIn | GitHub

PROFESSIONAL SUMMARY

Full stack AI engineer with 2+ years shipping production agentic, RAG and document-AI systems for banking, SaaS, retail, travel and EdTech clients. Own architecture across retrieval, inference and multi-tenant trade-offs, from Next.js frontends to FastAPI services and GPU inference, with measurable gains in accuracy, latency and reliability, plus client-facing scoping and mentorship.

TECHNICAL SKILLS

Agents & RAG: LangGraph, LangChain, Tool Calling, Hybrid Retrieval & Reranking, Embeddings, LLM Evaluation, Guardrails, Prompt Engineering

Models & ML: OpenAI API, LiteLLM, Meta Llama, Hugging Face, PyTorch, OCR & Document AI

Backend: Python, FastAPI, Node.js, Express.js, REST APIs, SSE Streaming, OAuth2 / JWT

Frontend: React.js, Next.js, JavaScript, TypeScript, Vite

Data & Vector Search: PostgreSQL, Qdrant, Redis, MongoDB, Supabase

Production & Cloud: AWS, Docker, GPU Inference, Langfuse, Nginx, CI/CD

PROFESSIONAL EXPERIENCE

Credentech Solutions | Software Engineer (AI) *Apr 2026 – Present*

- Own architecture decisions across retrieval, inference and multi-tenant trade-offs; standardized design patterns now used across **2+ production services**.
- Build full-stack AI products end to end: Next.js and React frontends, FastAPI services and REST APIs, PostgreSQL data models, and **RAG pipelines with evaluation and guardrails**.
- Design and prototype **agentic workflows** (LangChain, LangGraph, tool calling) and evaluate production feasibility ahead of platform roadmap decisions.
- Lead client-facing technical scoping for AI and full-stack deployments, and mentor an incoming engineer on RAG pipeline design and production debugging.

Credentech Solutions | Junior Software Engineer *Feb 2025 – Apr 2026*

- Delivered **4 client projects** across banking, SaaS, food and beverage, and travel, covering frontend, backend, AI services and deployment; progressed to **independent end-to-end ownership**.
- Built React and Next.js frontends and FastAPI and Node.js/Express backends on PostgreSQL.
- Developed OCR and document-intelligence services with LLM-based extraction, validation and structured JSON output, including **GPU inference with automated failover**.
- Secured APIs with OAuth2 and JWT, and built a PostgreSQL audit trail that reduced log-investigation effort by **~40%**.

Quelit | Data Science Intern *Jul 2024 – Dec 2024*

- Built a **persona-consistent chatbot** (OpenAI API, Streamlit, MongoDB) with conversation storage and analytics logging for 20+ users.
- Built an end-to-end **regression pipeline** in Python (EDA, validation, encoding, preprocessing) predicting student academic performance, comparing Scikit-learn models on R^2, RMSE and MAE; made ingestion, training and prediction pipelines reusable.

PROJECTS

Credentra Agents: AI Agent Builder Platform (B2B SaaS) *Credentech Solutions | 2026*

Next.js, FastAPI, PostgreSQL, Qdrant, LiteLLM, NeMo Guardrails, Langfuse, Docker

- Led engineering of a multi-channel agent builder (website widget, hosted chat page, WhatsApp) that runs every channel through one conversation runtime.

- Reached **~90% grounded-answer accuracy at ~2s latency** (1.38s median to first words) with four-branch Qdrant retrieval and a cross-encoder reranker.
- Raised retrieval **hit@1 from 0.77 to 0.95**; every answer cited its expected source and every unanswerable question was declined.
- Cut unsafe or off-policy responses by **~30%** via an 8-step model gateway (rate limits, NeMo guardrails, fallbacks, tracing) and a citation fence that fails closed.
- Cut new-business setup from **days to ~1 hour** with self-serve crawling and uploads; sped up critical-issue response **~40%** and content-gap discovery **~50%** with human takeover and a gap report.
- Isolated tenant data with PostgreSQL row-level security and workspace-scoped vector search.

OCR & KYC Pipeline: APB Bank, Laos *Credentech Solutions | 2025*

Python, FastAPI, Surya OCR, NVIDIA L4 GPU, PostgreSQL, Nginx, OAuth2 / JWT

- Built end to end an OCR microservice turning Lao passports, national IDs and family books into structured JSON inside the bank's digital onboarding flow.
- Reached **~95% field-extraction accuracy** (per field, 70+ samples per document type) at **500+ KYC documents a day**.
- Sustained **~99% uptime** with automatic **GPU-to-CPU failover** across two nodes on the bank's own servers.
- Secured the licensed B2B API with client credentials, short-lived JWTs, per-client licence expiry and a full audit trail.

Invoice Automation: Taco Bell Thailand (Siam Taco Co., Ltd.) *Credentech Solutions | 2026*

Python, FastAPI, Typhoon OCR, vLLM, React, Vite, PostgreSQL, Docker, Nginx

- Led an on-premises invoice processing app for mixed Thai/English supplier invoices across **5 supplier templates**.
- Hit **96% field accuracy** on 121 hand-checked invoices (exact match) with **0 silent errors**, flagging only 2.4% of fields for review.
- Replaced an LLM structuring step that invented values with OCR plus rule-based parsers; self-hosted the 2B OCR model at ~25s per invoice and ~200 an hour.
- Kept financial data on the client's own server, with a review dashboard and Excel, CSV and JSON export.

TourExc: Travel Booking Platform, Thailand *Credentech Solutions | 2026*

Next.js, React, TypeScript, Node.js, Express.js, PostgreSQL

- Rebuilt and rebranded a travel agency's platform (live at tourexc.com): website, partner-agent portal and admin panel on one backend.
- Shipped one cart for tours, activities, hotels and airport taxis, priced in the visitor's currency with online payment and booking confirmations (PDF invoice) by email and WhatsApp.
- Reached Lighthouse **SEO 92, accessibility 96 and best practices 96**.

AI-Powered CBSE Learning Platform *Freelance | 2026*

FastAPI, React, Next.js, Supabase, RAG, LLMs, Render

- Shipped a **2-month MVP** end to end (backend, AI services, frontend, deployment on Render), iterating on live demo-user feedback.
- Built curriculum-grounded RAG subject tutors personalised per class and subject.
- Designed an assessment engine tracking **25 CBSE-aligned parameters** (Bloom's Taxonomy, metacognition) with rule-based and LLM-assisted scoring.
- Cut estimated manual grading effort by **~50%** with an OCR and LLM answer-sheet grading workflow.

EDUCATION

Vellore Institute of Technology, Vellore, India *Sept 2020 – May 2025*

Integrated M.Sc., Computational Statistics and Data Analytics | CGPA: 7.64