SHEMAYON SOLOMAN

Machine Learning Engineer | AI Engineer
Dubai, UAE | +971 502531425 | shemayons@gmail.com
https://linkedin.com/in/shemayon-soloman | https://github.com/shemayon | https://shemayonsoloman.com

PROFESSIONAL SUMMARY

Machine Learning Engineer with 4+ years of professional experience in Python development, data engineering, machine learning, and production AI systems. Experienced in designing, building, deploying, and monitoring production AI systems using LLMs, Multi-Agent Systems, Multimodal RAG, Computer Vision, and MLOps on AWS and GCP across legal intelligence, enterprise analytics, healthcare, and generative AI.

CORE EXPERTISE

Generative AI & Agentic Systems: LLMs (GPT, Claude, Gemini), Prompt Engineering, Agentic AI, LangGraph, LangChain, MCP, Multi-Agent Systems, CrewAI, Stable Diffusion XL (SDXL)

LLM Evaluation & Retrieval Systems: RAG, CRAG, RAG Evaluation, Retrieval Optimization, Hybrid Search, Hallucination Reduction, Prompt Evaluation, Semantic Search, Grounded Responses, Citation-Based Retrieval, Vector Databases

Machine Learning & Deep Learning: Classification, Regression, Clustering, Recommendation Systems, Explainable AI (SHAP), Feature Engineering, Statistical Modeling, Deep Learning, Transformer Architectures, Self-Attention, Fine-Tuning (LoRA, PEFT), Hugging Face Transformers, Sentence Transformers, Quantization, Distillation, Diffusers

Computer Vision & Multimodal AI: YOLOv5, YOLOv8, U-Net, OpenCV, OCR Pipelines, Medical Imaging, Histopathology, Whole Slide Imaging (WSI), Disease Screening, Vision Language Models (VLMs), ControlNet

MLOps & Deployment: FastAPI, Celery, RabbitMQ, Docker, Kubernetes, MLflow, GitHub Actions, CI/CD, LangSmith, LangFuse, Model Monitoring

Programming, Cloud & Data Platforms: Python, SQL, PyTorch, TensorFlow, AWS (EC2, S3, Lambda, RDS, EKS, SageMaker), GCP (Vertex AI, BigQuery, Cloud Run, GKE), PostgreSQL, Redis, FAISS, Pinecone, Weaviate, Qdrant, Elasticsearch

PROFESSIONAL EXPERIENCE

Machine Learning Engineer

Mindsmap AI Services | Remote | Dec 2024 – Present

Legal Intelligence Multi-Tenant SaaS

- Architected and built a production-grade **multi-tenant Agentic Legal Intelligence SaaS** platform using **LangGraph**, enabling secure legal document intelligence and citation-grounded Q&A across **50K+ legal documents across multiple law firms** with organization-scoped retrieval.
- Engineered a **Multimodal Corrective RAG (CRAG)** pipeline integrating OCR, semantic chunking, metadata enrichment, hybrid retrieval, reranking, vector indexing, and citation-grounded generation, achieving **CER <2% (~98% key-field extraction accuracy)** on client-validated legal documents.
- Implemented **enterprise-grade tenant isolation** using RBAC, organization-scoped authentication, isolated vector indexes, metadata-filtered retrieval, and answer-level grounding to prevent cross-tenant data access.
- Automated CI/CD with **GitHub Actions** and deployed containerized services on **Google Cloud Run**; established **LangSmith** evaluation dashboards to monitor retrieval relevance, answer faithfulness, citation quality, and production system health.

Agentic Business Intelligence & Forecasting System

- Built a production-grade **multi-agent Business Intelligence & Forecasting** system using **LangGraph**, orchestrating specialized SQL, forecasting, visualization, and insight-generation agents to enable conversational analytics and self-service business intelligence for enterprise stakeholders.
- Engineered automated forecasting pipelines using **BigQuery ML (ARIMA & ARIMA_PLUS)** by integrating **ServiceTitan** operational and **Google Ads** marketing datasets, supporting model training, evaluation, retraining, and forecasting across **20+ business KPIs** while achieving **~10–12% MAPE** for selected monthly business forecasting tasks.
- Replaced legacy forecasting workflows with production ML pipelines, improving forecast consistency and enabling automated KPI reporting and AI-driven business insights.
- Automated CI/CD with **GitHub Actions** and deployed containerized services on **Google Cloud Run**; implemented model lifecycle management, forecasting performance tracking, data drift, concept drift detection, and production monitoring for reliable ML operations.

Multi-Agentic Job Search & Recommendation System

- Built a production-grade **multi-agent Job Search & Recommendation System** using **LangGraph** and **MCP**, orchestrating **5 AI agents** for intent understanding, query transformation, job search, recommendation, and conversational assistance.
- Developed an **LLM-powered query transformation service** that converted natural-language requests into structured search parameters, integrating with internal job search APIs to retrieve and rank relevant opportunities across **10K+ job listings**. Integrated real-time **voice AI** services using **OpenAI Realtime API**, **STT**, and **TTS**, enabling conversational job search and career assistance.
- Automated CI/CD with **GitHub Actions** and deployed containerized services on **AWS EC2**, using **Amazon S3** for artifact storage and metadata, with centralized logging and production monitoring.

Generative AI & Image Generation Pipeline

- Built an end-to-end **Generative AI platform** for domain-specific sports image generation, automating dataset creation from user-uploaded videos through frame extraction, **BiRefNet** background removal, **BLIP** caption generation, and SDXL dataset preparation.
- Engineered automated **LoRA fine-tuning** workflows using **Kohya-SS**, **Hugging Face Accelerate**, and **Stable Diffusion XL (SDXL)**, enabling GPU-accelerated model training, checkpoint versioning, and customized image generation.
- Deployed the platform on **AWS** using **FastAPI**, **Celery**, **RabbitMQ**, **Redis**, **PostgreSQL**, **Docker**, **Amazon S3**, **Amazon ECR**, and **GitHub Actions CI/CD**, implementing scalable asynchronous training, inference, and cost-efficient MLOps workflows.

Provided technical leadership by presenting solution architectures, leading client technical discussions, mentoring a team of 8 interns, conducting code reviews, and driving technical implementation under engineering management.

AI Engineer

CDAC – Centre for Development of Advanced Computing | Kerala, India | Mar 2024 – Dec 2024

Healthcare Intelligence & Digital Pathology System

- Built a production-grade **Healthcare AI System** integrating **digital pathology, medical computer vision, multilingual healthcare AI, and clinical decision support** to enable AI-assisted disease screening, diagnostics, and healthcare accessibility across multiple clinical workflows.
- Engineered an end-to-end **Whole Slide Image (WSI)** pathology pipeline for **Triple-Negative Breast Cancer (TNBC)** prognosis using **Efficient-UNet, U-Net, and YOLOv8**, integrating tumor-stroma segmentation, Tumor-Infiltrating Lymphocyte (TIL) detection, automated TIL scoring, and survival prediction, achieving a **Weighted Dice Score of 0.791**, **FROC of 0.572**, and **C-Index of 0.719**.
- Developed AI-assisted screening pipelines for **cervical cancer, oral cancer, and diabetic retinopathy** using **YOLOv5**, **YOLOv8**, and **OpenCV**, while building multilingual healthcare workflows integrating **OCR**, **Speech-to-Text**, **translation**, and **Vision LLMs** for conversational access to medical reports and clinical documents.
- Contributed to the **United Nations Inter-Agency Task Force Award-winning "Digitally Connected Tribal Colonies"** initiative by delivering AI-powered healthcare screening solutions for non-communicable disease programs and supporting scalable clinical deployment workflows.

Data Analytics Engineer (Python & Machine Learning)

MattGloss Advertising | Kerala, India | Jul 2020 – Jun 2022

- Built **Python** and **SQL** ETL pipelines for data ingestion, transformation, and reporting, while managing large-scale datasets and maintaining **PostgreSQL** databases to support analytics and downstream machine learning workflows.
- Automated reporting workflows and developed interactive **Tableau** and **Power BI** dashboards for KPI monitoring, campaign performance analysis, customer engagement, and business decision-making in collaboration with cross-functional stakeholders.
- Developed a **customer churn prediction** model for an e-commerce client using feature engineering and machine learning techniques to identify at-risk customers, supporting targeted retention strategies and model experimentation.

AWARDS

United Nations Inter-Agency Task Force Award (2024)

Recognised for contributions to the Digitally Connected Tribal Colonies initiative, delivering AI-powered healthcare screening systems supporting underserved tribal communities across multiple non-communicable disease programs in India.

EDUCATION

M.Sc. Computer Science and Engineering (specialization in Machine Intelligence)

Kerala University of Digital Sciences, Innovation and Technology (DUK) | 2022 – 2024