

Full-stack Machine Learning Engineer who independently architects and delivers complete AI products—building scalable ML pipelines, backend infrastructure, and frontend applications for production systems in LLMs, RAG, and Computer Vision.

 joabdavid96@gmail.com

 +49 1742004685

 github.com/Joab-David-Johanan

PORTFOLIO PROJECTS

Multi-Agent AI Application

- Designed and deployed a configurable multi-agent LLM system enabling dynamic behavior selection and tool-based reasoning using LangGraph.
- Reduced inference latency by 80% by implementing multi-layer caching (session cache and Redis-backed global cache) and streaming response generation for faster perceived response times.
- Improved user engagement by 50%. Integrated conversational memory and dynamic follow-up suggestion generation to enable context-aware multi-turn conversations.
- Optimized system performance by up to 40%. Caching LLM instances, agent initialization workflows, and external tool results helped minimize repeated API calls, lowering operational cost and improving latency.
- Integrated real-time observability dashboard for monitoring API calls, latency, and errors using Prometheus and Grafana.
- Engineered a modular multi-agent architecture with domain-specialized assistants (general, financial, medical, legal) to improve response reliability and enforce clear separation of concerns for domain-specific queries.
- Containerized the application using Docker and deployed it on AWS Fargate with automated CI/CD pipelines via Jenkins. Implemented logging, monitoring, and code quality validation using SonarQube.
- Tech Stack:** Python, LangChain, LangGraph, FastAPI, Streamlit, Redis, Docker, Jenkins, OpenAI API, Groq API, AWS (ECR, Fargate)

RAG Agent Framework

- Built a modular Retrieval-Augmented Generation (RAG) research framework to experiment with multiple RAG architectures including Adaptive, Corrective, Multi-Agent, Multi-Modal, and Cache-based retrieval strategies.
- Designed a document ingestion pipeline that converts uploaded text files into vector embeddings using FAISS for semantic search and retrieval.
- Implemented a FastAPI backend that handles file ingestion, vector store creation, and routing of queries to different RAG pipelines.
- Developed a Next.js + React frontend with an interactive chat interface and sidebar-based controls for selecting different RAG strategies.
- Implemented a multi-agent workflow using LangGraph where specialized agents collaborate to retrieve information and generate responses.
- Tech Stack:** Python, FastAPI, LangChain, LangGraph, FAISS, HuggingFace Embeddings, Groq LLMs, Next.js, React, TailwindCSS

YOLO-based Image Search System

- Built an interactive computer vision search application that performs object detection on image collections using multiple YOLO model variants (YOLOv11 and YOLOv26).
- Generated structured metadata from detection results to enable class-based image retrieval and visual exploration.
- Designed an inference pipeline that processes image directories and stores detection outputs (bounding boxes, classes, counts) as metadata for efficient downstream querying.
- Implemented support for multiple model weights to compare performance across lightweight and high-capacity YOLO variants.
- Developed a Streamlit-based interface enabling users to filter images using logical search conditions (AND / OR) and configurable class-count thresholds, allowing flexible exploration of detected objects within large image collections.
- Implemented dynamic visualization of results with bounding box overlays, grid-based display controls, and metadata highlighting to improve interpretability of model predictions.
- Tech Stack:** Python, PyTorch, YOLOv11, YOLOv26, Streamlit, FastAPI, Computer Vision Pipelines

TECHNICAL EXPERTISE

Programming Languages:	Python, JavaScript (ES6+), SQL	Machine Learning Frameworks:	PyTorch, TensorFlow, Keras, Scikit-learn, AutoML
LLM & Generative AI:	LangChain, LangGraph, Retrieval-Augmented Generation (RAG) pipelines, Prompt engineering & evaluation	Frontend Technologies:	React, Next.js, HTML5, CSS3
Backend & API Development:	FastAPI, Django, Streamlit	Cloud-based Technologies:	AWS - S3, EC2, ECR, SageMaker, GroundTruth, Fargate, Lambda



PROFESSIONAL CERTIFICATIONS

AWS Machine Learning Specialty 
Amazon Web Services Certification

AWS Machine Learning Associate 
Amazon Web Services Certification



WORK EXPERIENCE

Visual Inspection Expert InspectifAI GmbH

Karlsruhe, Germany

InspectifAI helps in the deployment of AI solutions for visual inspection tasks with the mission to improve drug & patient safety.

Achievements/Tasks

- Contributed to training and validation of computer vision models for pharmaceutical vial inspection.
- Improved model performance by enhancing dataset quality and consistency across large-scale image datasets (10k+ images/week)
- Optimized model evaluation and data preprocessing workflows using Amazon SageMaker.
- Improved dataset consistency and annotation quality for production ML deployments.
- **Technologies:** Python, AWS SageMaker, AWS GroundTruth, Computer Vision Pipelines

Contact: Süleyman Salihi (Project Manager) – Sueleyman.Salihi@inspectif.ai

Internship DATEV eG

Nuremberg, Germany

Datev eG is a software house and IT service provider for tax consultants, auditors and lawyers.

Achievements/Tasks

- Developed a Django-based web annotation tool to support OCR model training for tax document processing.
- Designed a PostgreSQL database schema for structured annotation storage.
- Improved labeling workflow efficiency for document classification and OCR pipelines.
- Documented system architecture and usage guidelines for internal ML teams.
- **Technologies:** Python, Django, PostgreSQL

Contact: Neha Pawar (Data Scientist) – epochs19@gmail.com



EDUCATION

Masters (M.Sc.) in Computational Engineering Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

Notable Courses :

- Deep learning, Pattern recognition, Pattern analysis, Image processing, Machine learning, Numerical simulation, Computer aided engineering, Advanced programming techniques.
- Master thesis in "Development and Evaluation of a Modular Machine Learning Pipeline for Monitoring Manufacturing Processes Based on Force, Torque or Sound Signals".

Bachelors (B.Tech) in Mechanical Engineering Sharda University, India

Notable Courses :

- Object oriented programming, Robotics engineering, Advanced manufacturing processes, Aerospace vehicles, Fluid dynamics, Thermodynamics, Material science.
- Bachelor thesis in "Extraction of Biodiesel from Canola Oil".



LANGUAGES

English – C1
Full Professional Proficiency

German – B1
Limited Working Proficiency

Hindi – C2
Native or Bilingual Proficiency

Tamil – C2
Native or Bilingual Proficiency