BHUVNESH KUMAR

Technical AI Team Lead | iambhuvi7@gmail.com | LinkedIn |+91-7417367313| Bengaluru, India

Summary

I'm a Technical AI Team Lead with 8+ years across GenAI agents, computer vision, and MLOps. Currently leading AI agent platforms for Cisco optical networking — shipping the first MCP-protocol AI tool server in CONC 26.1.1. Day-to-day I build with LangChain, LangGraph, AutoGen, Pydantic AI, and OpenAI-compatible APIs across LLM providers — shipped on AWS and Docker.

Skills

- **Autonomous AI Agents** (LangGraph, AutoGen, Multi-Agent Systems)
- **Advanced RAG Architectures** (LlamaIndex, Contextual Retrieval, Vector DBs)
- **LLM Fine-Tuning & SFT** (Supervised Fine-Tuning, ChatML, Transformers)
- **Computer Vision** (YOLOv5, Semantic Segmentation, OCR/TrOCR)
- **MLOps & Model Serving** (GPU Optimization, Inference Scaling)
- **Kubernetes & Cloud Native** (AWS EKS, Docker, Helm)
- **System Design for AI** (Microservices, Event-Driven Architecture)
- **API Development** (FastAPI, AsyncIO, REST)
- **Observability** (EFK Stack, Log Analysis, Automated RCA)
- **Technical Leadership** (Team Mentoring, Agile, Architecture Review)

Experience

Technical Team Lead (GenAI & Computer Vision)

CLIENT CISCO, VCTI, Bangalore *Nov-2022*

- **CONC MCP Server (26.1.1):** Led delivery of the first MCP-protocol AI tool server for Cisco optical networking — positioned as a standard microservice in the NxF framework.
- **AINetOps Beta:** Owned end-to-end development and early field-trial delivery (FastAPI, LangGraph, React, MCP, Docker).
- **Deep Agent for Automated RCA:** Architected a "Deep Agent" workflow using LangGraph to autonomously investigate customer-reported incidents. The agent traverses system logs and process flows to generate L1 triage reports and pinpoint root causes, accelerating developer workflows and reducing Mean Time to Resolution (MTTR) by 50%.
- **Agentic AI Architecture:** Architected an autonomous multi-agent system using AutoGen and LangGraph to automate technical documentation generation from meeting transcripts, reducing manual effort by 90%.
- **Production RAG Systems:** Spearheaded the development of a Contextual RAG pipeline using LlamaIndex and private enterprise data. Achieved 80% Context Recall by implementing advanced retrieval strategies and custom embedding fine-tuning.

- **MLOps & Scalability:** Led the migration of inference pipelines to AWS EKS (Kubernetes). Implemented GPU optimization strategies that reduced inference latency and cloud costs.
- **Geospatial Computer Vision:** Improved pole detection accuracy by 48% by designing a semantic segmentation model (DeepLabv3+) for satellite imagery to estimate road width and classify infrastructure.
- **Engineering Leadership:** Established the EFK stack for observability, standardized Docker containerization workflows for faster CI/CD, and mentored the team on best practices for ChatML data formatting and SFT (Supervised Fine-Tuning).

Senior Software Engineer (Computer Vision)

CLIENT HSBC BANK, Wipro *Oct 2021 – Nov 2022*

- **High-Precision Object Detection:** Engineered an ID-card localization system using YOLOv5, achieving a state-of-the-art 95% mAP in production environments.
- **Multimodal Document Understanding:** Designed a hybrid classification engine combining Visual features (VGG16) and Textual features (TF-IDF), boosting document classification accuracy from 70% to 95%.
- **Inference Optimization:** Conducted deep profiling of Python pipelines, refactoring bottlenecks to achieve a 4x reduction in request latency and significant memory optimization.
- **OCR Solutions:** Deployed Transformer-based OCR models (TrOCR, PaddleOCR) for extracting key-value pairs from multi-lingual handwritten documents with 90% accuracy.

Senior Software Engineer (Computer Vision)

ARTIVATIC.AI, Bangalore *Jun 2021 – Oct 2022*

- **Identity Verification Pipeline:** Developed a complete e-KYC solution using Siamese Networks for face matching and dlib for liveness detection, critical for insurance fraud prevention.
- **Predictive Modelling:** Built and fine-tuned classifiers to predict customer risk attributes (Age, Gender, BMI, Smoker Status) from facial imagery.
- **Document Intelligence:** Improved signature fraud detection accuracy from 45% to 81% by implementing a ResNet50V2 backbone for signature matching.

Full Stack Developer

SNAP-ON BUSINESS SOLUTIONS, Noida *Jun 2018 – Jun 2021*

- Developed an automated Chassis number detection system utilizing Image Processing techniques and OpenCV.
- Built scalable web applications using Angular and Java, adhering to Agile methodologies and MVC design patterns.

Certifications

- **Building deep agents with LangGrap [2025]:**Certificate Link
- **Search in the LLM era for AI engineers [2024]:** Expertise in customizing search backends to address industry- specific query requirements. Certificate link
- **DeepLearning.ai on AI agents [2024]:** Build agentic AI workflows using LangChain's LangGraph and Tavily's agentic search. Certificate link

Education

- Computer Science and Engineering, B.Tech [2016] - Uttarakhand Technical University, Dehradun, Uttarakhand
- Senior Secondary School, CBSE [2012] - DAV Centenary Public School, Haridwar
- High School, CBSE [2010]- DAV Centenary Public School, Haridwar