

SUMMARY

- Lead Engineer with 5 years of hands-on experience in cutting-edge technologies, proficient in developing and deploying GenAI, RAG, Agentic AI, and deep learning models from scratch.
- Built a production-grade AI-powered a platform using FastAPI microservices, OpenAI Realtime API, Twilio, Redis, MongoDB, and Milvus RAG — enabling real-time voice AI, live human handoff, and enterprise knowledge retrieval at scale.
- Built a production-grade RAG platform for SEBI legal & regulatory documents — combining hybrid semantic + BM25 search, reranking, citation-constrained LLM generation, Redis caching, and secure offline Docker deployment for compliance and legal teams.
- Recognized for significant contributions to research and development, including the publication of two papers at IEEE International conferences (awarded one Best Paper).
- Expertise spans advanced domains including NLP, Autonomous Driving, Computer Vision, and Camera-LiDAR-Radar Fusion.
- Highly skilled in architecting machine learning and deep learning models, modifying neural network architectures, and optimizing data processing pipelines.

EDUCATION

Indian Institute Of Technology (IIT) Bhilai
BTech in Electrical Engineering

Chhattisgarh, India
July 2017 – July 2021

TECHNICAL SKILLS

Languages: Python, C/C++

Developer Tools: GitHub, VS Code, Linux, Docker, GDB/PDB

Libraries: Pandas, NumPy, Matplotlib, Scikit-learn, PyTorch, TensorFlow, OpenCV, Hugging Face, Google ADK

Other: Generative AI, LLMs, Computer Vision, ADAS, Camera-LiDAR-Radar Fusion, Data Structures, Agent, Rag, vLLM, VectorDB, Redis, SmLLM, OCR, Google Cloud, Agent Engine

EXPERIENCE

HCL Technologies Limited
Lead Engineer

Noida, India
June 2025 – Present

- IFRS

- Developed an AI-powered Invoice Anomaly Detection Agent (Gemini/Google ADK) that analyzes invoice risk using vendor history, confidence scoring, and evidence-backed signals to enable automated severity-based routing.
- Developed a Google ADK-powered FP&A Variance Analysis Agent that automates SAP actuals vs budget/forecast/prior-period analysis, validates data, identifies root-cause drivers, and generates management-ready variance commentary and exception reports.
- Ensured production readiness by deploying both agents on Vertex AI, backed by PostgreSQL/Cloud SQL for data persistence and monitored with Prometheus observability.
- **Tech Stack:** Python, GCP, GenAI, LLMs, Agents, Docker, FastAPI, Cloud SQL, Google ADK, Vertex AI, Agent Engine, Cloud Run.

- AACA

- Architected a multi-service AI contact-center platform (AACA) using a microservices design pattern — separating a latency-sensitive realtime voice edge (FastAPI + WebSocket) from a stateful business control plane (FastAPI + MongoDB + JWT auth), enabling independent scaling, deployment, and team ownership across each domain.
- Built a bidirectional voice streaming service bridging Twilio Media Streams with the OpenAI Realtime API over WebSockets streaming native g711_ulaw audio to minimize transcoding overhead, speech-to-speech response, achieving sub-second voice AI latency.
- Implemented a Retrieval-Augmented Generation (RAG) pipeline using OpenAI embeddings + Milvus vector search for enterprise knowledge-base retrieval, with structured GenAI tool calling (kb_search, customer_verify, ticket workflows) enforcing identity verification before sensitive operations — deployed as real-time callable tools inside live voice sessions.
- Designed a horizontally scalable Redis pub/sub event bus enabling cross-instance WebSocket delivery — HTTP routes publish to **agent_events**:id channels; a WebSocket hub on any backend instance delivers events to all open agent browser tabs, supporting live call assignment, interrupt/resume/transfer signals, and AI handoff coordination without inter-service direct coupling.
- Containerized all services using Docker, orchestrated local infrastructure (MongoDB, Redis, Milvus, MinIO, PostgreSQL) via Docker Compose, with architecture designed for Cloud Run horizontal scaling — stateless voice edge workers, Redis-backed state.
- **Tech Stack:** Python, GCP, GenAI, LLMs, Agents, RAG, Cloud Run, Twilio, Docker, FastAPI, OpenAI Realtime API, WebSockets, Redis, MongoDB, Milvus, JWT

- SEBiSearch

- Designed a full Retrieval-Augmented Generation pipeline — PDF parsing (Docling) → heading-aware chunking → Jina v5 embeddings → Milvus vector index → hybrid retrieval → reranking → citation-constrained LLM answer generation over thousands of SEBI regulatory documents.
- Engineered a hybrid retrieval pipeline (Jina v5, BM25) with ModernBERT cross-encoder reranking, achieving an 80% retrieval hit rate and 75.8% nDCG.
- Mitigated LLM hallucinations by developing citation-constrained generation (vLLM/OpenAI), grounding all answers with exact file, page, and order metadata.
- Optimized system latency via multi-layered caching (Redis, FAISS) and integrated Prometheus metrics and health probes, concurrency controls, and request timeouts for production-grade reliability.
- Designed 100% offline, internet-free deployments for restricted SEBI environments via Docker Compose service with RHEL 9 / air-gapped VM support.
- **Tech Stack:** Python, FastAPI, Milvus, Jina v5, ModernBERT, vLLM/OpenAI, Redis, FAISS, Docker, Docling, Embedding, BM25, Reranker, hybride-search.

PathPartner Technology (member of KPIT)

Senior Software Engineer

Bangalore, India

July 2021 – May 2025

- Image Stitching

- Transformed feature detection deep learning model into ONNX and target device formats for seamless integration into C++ pipeline.
- Created a C++ pipeline to incorporate a deep learning feature detection model and matching algorithm.
- Deployed the pipeline on a Qualcomm device for real-time edge processing, reducing CPU runtime by 75%.
- **Tech Stack:** Python, C++, OOP, PDB Debugger, NumPy, PyTorch, VS Code, Linux, QNN

- CES

- Effectively trained SOLOv2 on customized datasets, showcasing proficiency in cutting-edge technologies.
- Implemented OpenCV-based image blending, contributing to enhanced visual outputs of BEV Images.
- Transformed an object instance segmentation model into a multi-model architecture and integrated a multi-image classification network. Developed and executed end-to-end train/test pipelines.
- Utilized LIME for precise feature extraction from the YOLO model and subsequent visualization.
- **Tech Stack:** Python, OOP, PDB Debugger, NumPy, PyTorch, mmdet, OpenCV, VS Code, Linux

- Deep Learning Kernels

- Developed a Software Suite integrated with PyTorch and TensorFlow to facilitate high performance and efficiency of deep learning training & inference on accelerators.
- Analyzed and resolved kernel bugs, successfully mapping PyTorch kernels to TPC kernels in C++.
- Developed operators using combinations of TPC kernels and converted deep learning models to ONNX, visualizing them using the Netron app.
- **Tech Stack:** Python, C++, NumPy, Pandas, PyTorch, TensorFlow, VS Code, Linux, ONNX, Git, JIRA

AgVerse Technology Pvt. Ltd

Software Engineer Intern

Bangalore, India

Feb 2020 – Jun 2020

- Contributed to bringing up Firebase server support to pull and analyze telemetry data.
- Created a pipeline mechanism where data preprocessing modules can be integrated on demand.
- Performed Baum-Welch algorithm-based parameter and state estimation.
- **Tech Stack:** Python, Firebase, HMM (Hidden Markov Model), Pandas

PUBLICATIONS

“BEV-IMTP : Lightweight multi task learning for joint birds eye view instance mapping and trajectory prediction” (Aug. 2026).
In: *Discover Artificial Intelligence*. URL: <https://doi.org/10.1007/s44163-026-01970-1>.

“ClothFormer - A Boundary Aware Self-attention Network for Human Outfit Parsing” (Nov. 2023). In: *ICRAIS 2023*. URL: <https://ieeexplore.ieee.org/document/10367184>.

“DCSP-RCNN: A Network for Person Detection with Color, Size and Pattern Characteristic Parsing” (Nov. 2023). In: *ICRAIS 2023*. URL: <https://ieeexplore.ieee.org/document/10367168>.