

ZEESHAN SHAIKH

📍 Pune, Maharashtra, India 📩 zinczee97@gmail.com ☎ 9890895611 💬 in/zeezinc 🌐 zeezinc.github.io

SUMMARY

Dynamic Gen AI Data Specialist with ~4 years of experience in integrating large language models with vector-based RAG pipelines and computer-vision-based surveillance platforms for enterprise analytics. Architected AI solutions that decreased manual effort by up to 70% and boosted system performance threefold with Python, FastAPI, and cloud infrastructure.

EXPERIENCE

Gen AI Data Enthusiast

Hexalytics

June 2024 - Present, Pune, India

- Developed Generative AI chatbots and analytics solutions by integrating large language models (LLMs) with vector-based RAG pipelines and multi-step AI workflows to support enterprise data needs, handling huge number of document queries per month.
- Built and delivered AI-powered platforms for education and data analytics reducing manual effort by 50–70% and improving system performance by up to 3x through caching and backend optimization.
- Designed and maintained Python-based backend systems using FastAPI to automate document ingestion, search, summarization, and report generation across 10k+ unstructured documents.
- Implemented end-to-end document AI pipelines using Python, LlamaParse, Whisper, OCR tools, metadata-aware chunking, and embeddings, enabling efficient processing of PDFs, Word, PPT, CSV, URLs, audio, and video.
- Applied LLMs for natural language interfaces, including chatbots and Natural Language to SQL insight generation, reducing analyst dependency by ~50% and improving data-driven decision-making.

Backend Data Engineer

HftSolution

October 2021 - February 2024, Pune, India

- Designed and delivered production-grade Python backends (FastAPI, Flask) powering AI-driven search and surveillance systems, supporting real-time querying, analytics, and ML inference workflows.
- Built a traffic surveillance and violation detection system using computer vision and deep learning, implementing vehicle counting, helmet detection, triple-riding detection, and number-plate localization across 10,000+ images and video frames, enabling automated enforcement and traffic density insights.
- Developed a high-performance real-estate search backend with real-time query generation, heuristic-based ranking, and multi-layer caching, improving response times by ~3x and reducing backend load by ~40%.
- Implemented data preparation and analytics pipelines to transform raw search and surveillance data into ML-ready datasets, enabling traffic trend analysis, search behavior insights, and future model experimentation.
- Integrated rule-based heuristics, ML patterns, and anomaly detection, reducing manual effort by 60–70% while maintaining production stability.

Backend Intern

Transition Computing

December 2019 - June 2020, Pune, India

- Modernized legacy backend through service-layer refactoring, implementing layered architecture (controller, service, repository) patterns.
- Contributed to backend modules involving RESTful API development, focusing on request/response validation, data consistency, and clean API contracts.
- Collaborated with backend engineers to integrate ORM-based data access layers, improving code consistency, supporting automated database migrations, and reducing technical debt across data-centric modules.
- Followed Git-based collaborative workflows, including feature branching, pull requests, and structured code reviews in a production environment.

PROJECTS

AI-Powered RFP & Document Intelligence Suite

Hexalytics

- Built an end-to-end AI platform for processing and analyzing RFP / PDF / Word / PPT / CSV / URL / audio / video documents.
- Engineered advanced RAG/CAG pipelines, multi-model prompting, and structured extraction workflows.
- Implemented file ingestion and parsing using LlamaParse, OCR, PyMuPDF, Whisper (audio), and custom chunking logic.
- Automated TOC extraction, header-section mapping, metadata tagging, and python-docx report generation.
- Reduced manual Request for Proposal review time by 60% and improved extraction accuracy by ~40%.
- Tech Stack: Python, LangChain, LlamaParse, ChromaDB, Groq, Vertex AI, FastAPI, Docker.

Smart Teaching System

Hexalytics

- An intuitive generative AI system used by multiple Middle-East educational institutes.
- Built AI tools for lesson planning, assignments, MCQs, summaries, vocabulary generation, email writing, chapter insights, text-to-speech, and speech-to-text.
- Integrated fine-tuned LLMs for personalized academic content generation.
- Reduced teachers' manual workload by 70% and improved student experience with auto-generated learning materials.
- Tech Stack: OpenAI, Fine-Tuning, FastAPI, React, Python, Prompt Engineering.

Erdi Document Chatbot

Hexalytics

- Interactive conversational chatbot allowing users to ask questions over PDFs, Word, PPT, CSV, audio, video, and URLs.
- Built ingestion pipelines including OCR, file converters, markdown extraction, URL scrapers, and metadata-aware chunking.
- Implemented multi-document memory, follow-up reasoning, and collection versioning.
- Enabled dynamic context retention with high-accuracy vector retrieval.
- Tech Stack: LangChain, Groq, ChromaDB, FastAPI, React, LlamaParse, Whisper, Python.

Student Information Bot

Hexalytics

- AI agent that converts natural language into SQL queries and generates data-driven insights.
- Built NL → SQL → execution pipelines with an auto-correction layer to fix SQL syntax errors.
- Provided advanced analytics including gender analytics, enrollment trends, and school performance insights.
- Reduced data analyst dependency by 50% through automated query generation and insight delivery.
- Tech Stack: Gemini, FastAPI, LangGraph, Snowflake, Python, SQL.

Property Finder Dashboard – Real-Estate Search Platform

HftSolution

- Built a high-performance backend for a large real-estate search platform, supporting complex filters and high-traffic query workloads.
- Implemented dynamic query generation with rule-based ranking and heuristic scoring, using signals for example location, price, attributes, freshness, and user filters to improve search relevance.
- Designed Redis-based caching layers for ranked search results and frequent filter combinations, improving query performance by ~3x and reducing database load.
- Analyzed search and filter usage data to refine ranking heuristics and caching strategies, lowering Redshift analytics load by ~40% and stabilizing response times.
- Tech Stack: Python, FastAPI, Ranking & Heuristics, Redis, AWS (EC2, ECS, Secrets Manager, Redshift), Docker, Boto3, Async Jobs.

ThirdEye - Intelligent Traffic Surveillance System

HftSolution

- Designed and implemented a decoupled traffic monitoring platform using a Django frontend and Flask-based ML backend to analyze images and videos for traffic violations and traffic density monitoring.
- Built computer vision pipelines for vehicle counting using background subtraction and contour detection, helmet detection using MobileNet SSD combined with a custom Keras classifier, and triple riding detection using YOLOv3.
- Implemented number plate localization using Haar Cascade classifiers for vehicle identification workflows.
- Processed and analyzed 10,000+ images and video frames, enabling automated violation detection, helmet non-compliance identification, and traffic analytics to support road safety enforcement.
- Tech Stack: Computer Vision, Machine Learning, Python, Flask, Django, OpenCV, TensorFlow, Keras, YOLOv3, ImageAI, NumPy, Pandas.

EDUCATION

Master's in Science(Computer Science)

Savitribai Phule Pune University • Pune, India • 2020 • 8.25

Indian Certificate Of Secondary Education (ICSE)

The Bishop's School (CAMP) • Pune, India • 2013 • 87%

SKILLS

Programming & Backend: Python, FastAPI, Flask, Django, REST APIs

AI & Machine Learning: Large Language Models (LLMs), Generative AI, Retrieval-Augmented Generation (RAG) Pipelines, Prompt Engineering, LangChain, LlamaIndex, LangGraph, Transformers, Embeddings

Document & Data Processing: NLP, Document Parsing & Classification, Table & Metadata Extraction, File Conversion Pipelines, Image & Video Processing, OCR, Speech-to-Text, Text-to-Speech, Traffic & Surveillance Data Processing

Automation & Pipelines: Machine Learning, Deep Learning, Computer Vision, Object Detection, Model Training & Evaluation

Data Engineering & Analysis: Pandas, NumPy, Feature Engineering, Data Preprocessing for ML Readiness

Databases & Caching: PostgreSQL, MySQL, Redis, Snowflake, Vector Databases (ChromaDB, Qdrant)

DevOps : Docker, Git, Bitbucket, GitHub Actions, Jenkins, AWS (EC2, ECS, Secrets Manager, Redshift), CI/CD

Soft Skills: Problem-solving, Critical Thinking, Communication, Collaboration