

Gaurav Pandey

+91 7017176107 | gaurav1237890@gmail.com | linkedin.com/Gaurav Pandey | github.com/Gaurav Pandey | gaurav327.github.io/portfolio/

PROFESSIONAL SUMMARY

AI Engineer with 2+ years of hands-on experience building production-grade AI applications. Proven expertise in designing end-to-end AI solutions, including Retrieval-Augmented Generation (RAG), agentic workflows, and model fine-tuning. Delivered industry-impacting solutions achieving 98%+ accuracy, and secure enterprise self-service assistants. Strong background in Python, Flask, FastAPI, LangChain, LangGraph, CrewAI, agentic architectures, Generative AI (LLMs, RAG, Fine-Tuning), Databases, Docker with a consistent track record of transforming complex business problems into scalable and reliable AI products.

EDUCATION

Graphic Era Hill University	2019 – 2023
Bachelor of Technology in Computer Science and Engineering (CSE)	CGPA: 8.68
BLM Academy	2018 – 2019
Intermediate	Percentage: 63.6%
White Hall School	2016 – 2017
Matriculation	CGPA: 8.0

WORK EXPERIENCE

HL Mando Softech India	Gurugram
Engineer	July 2025 – Present
<ul style="list-style-type: none">Tech Stack: Python, Flask, FastAPI, Fine-Tuning, YOLOv11, DeepseekOCR, OpenCV, MySQL, ChromaDB, MinIO, LangChainDeveloped AI Discrepancy Highlighter to compare Master vs Test 2D engineering brake design drawings, using YOLOv11 to detect and localize dense parameter regions and DeepseekOCR to extract dimensions and tolerances, achieving 98%+ parameter recognition accuracy in a domain where no existing solution performs perfectly.Designed and built Vaani – Workforce Self-Service AI, a secure HR/IT policy assistant implementing authentication, data pipeline, ChromaDB semantic retrieval with reranker, MySQL guardrails and prompt/query safeguarding, and employee-wise audit logs for traceability.Collaborated with cross-team stakeholders to enable automated QA inspection and internal self-service intelligence, reducing manual validation effort and improving response reliability.	
Impressico Business Solutions	Noida
Trainee Engineer	January 2024 – June 2025
<ul style="list-style-type: none">Tech Stack: Python, Flask, OpenAI, PostgreSQL, PineconeDB, Langchain, Langgraph, CrewAI, Docker, GitHubDeveloped a Product Recommendation System with an electronics dataset stored in PostgreSQL, integrating SQL-based search and Semantic Search engines for data retrieval and passing results to an LLM for personalized recommendations.Designed and developed a chatbot using Generative AI, LLMs, RAG, OpenAI, LangChain, and HuggingFace, with a data pipeline to store, clean, preprocess, and generate embeddings for efficient query processing.	
LTI Mindtree Limited	Remote
Internship – Graduate Engineer Trainee (certificate)	February 2023 – May 2023
<ul style="list-style-type: none">Gained practical experience and insights into IT work and industry through the internship.Worked on Java and C# modules, including study materials and related tasks.Completed assigned activities on time based on study materials.	

TECHNICAL SKILLS

Programming: Python, Linux CLI	Frameworks: Langchain, Langfuse, Streamlit, Flask
Databases: MySQL, PostgreSQL, PineconeDB, ChromaDB	Libraries/Tools: Numpy, Pandas, GitHub, Label Studio
DevOps: AWS, Docker, Linux	Generative AI: LLMs, GPT Models, RAG, HuggingFace, OpenAI
AI/Agentic: LangGraph, CrewAI, Tool-Calling Agent Workflows	Model Fine-Tuning: Custom dataset creation, LLM Fine-Tuning

PROJECTS

AI Discrepancy Highlighter
<ul style="list-style-type: none">Tech Stack: Python, YOLOv11, DeepseekOCR, Label Studio, OpenCV, Image Preprocessing, Object Detection, Optical Character Recognition (OCR), Coordinate Mapping, VisualizationDesigned a Computer Vision + OCR pipeline to compare Master vs Test 2D engineering drawings of brake designs for automated quality validation.

- Annotated drawing sections to label **critical parameters (Dimensions, Tolerances, GD&T, etc.)** using **Label Studio** to build a custom engineering-OCR dataset.
- Trained **YOLOv11 object detection model** to precisely **detect and localize parameter regions** in dense CAD drawings — enabling region-based reading instead of full-image OCR.
- Fine-tuned and trained **DeepseekOCR** on YOLO-cropped regions to extract engineering parameters with high precision.
- Mapped OCR outputs using **coordinate alignment** and built a comparison algorithm to identify deviations between master and test drawing values.
- Addressed industry challenge: **No existing solution works perfectly on complex engineering drawings**, achieving a highly reliable **98%+ parameter recognition accuracy** for automated QA.

Vaani – Workforce Self-Service AI

- **Tech Stack:** Python, Flask, LangChain, ChromaDB, HuggingFace Transformers, Qwen AI, BGE Reranker, RAG, MySQL, MinIO (S3-compatible), SQL Guarding, Employee-wise Logs
- Engineered a secure **HR/IT policy knowledge system** enabling **self-service actions and intelligent retrieval** from handbooks.
- Implemented **authentication** and strict **prompt/query guardrails** to ensure **authorized and safe access**.
- Designed semantic **PDF chunking and embedding pipeline** for **robust policy recall without external assumptions**.
- Integrated **Chroma retriever + Reranker** to perform **ranked context selection in optimized batches** with limited dependencies.
- Ensured **policy-compliant retrieval and ranked context selection** as the core reliability driver of the assistant.
- Enforced strict **MySQL guardrails** for employee and form search, preventing unsafe query behavior.
- Delivered **personalized policy responses** using employee attributes (grade, team, designation).
- Enabled **employee-wise log storage** for complete **traceability of auth events, SQL outputs, fallbacks, and responses**.
- Integrated MinIO for **profile image upload/retrieval** and **form access through presigned URLs**.
- System behavior optimized for **security, relevance ranking, and internal self-service automation**.

Retail Concierge Services

- **Tech Stack:** Python, Flask, OpenAI GPT API, PostgreSQL, PGvector, LangChain, HuggingFace Transformers, Semantic Search, Sentiment Analysis, RAG
- Built an AI product concierge and recommendation engine for electronics dataset, **combining SQL + semantic intelligence** for precise suggestions.
- Collected, structured, and stored large-scale product catalog data in **PostgreSQL**, enabling fast indexed and filtered retrieval.
- Developed a **Flask REST API backend** to orchestrate query processing, database search, embeddings lookup, and LLM-powered recommendations.
- Implemented **hybrid search** using **SQL-based filtering** and **PGvector semantic similarity search** with metadata constraints.
- Used **SQLDatabaseChain (LangChain)** to transform natural queries into SQL, execute database lookup, and used semantic search for advanced filtering.
- Merged outputs from both search systems and fed them into an **OpenAI LLM** for ranked, user-aligned product recommendations.
- Integrated **sentiment analysis on product reviews** (positive/neutral/negative classification) to enhance personalization and recommendation relevance.
- Applied a full **Retrieval-Augmented Generation (RAG)** strategy to unify **structured + unstructured data** for richer recommendations.
- Improved decision reliability by reducing search noise and delivering concise, context-aware, review-grounded suggestions.

Impressico's Chatbot

- **Tech Stack:** Python, Flask, OpenAI GPT API, PostgreSQL, LangChain, HuggingFace Transformers, Vector Embeddings, RAG
- Developed an internal AI chatbot to deliver **personalized, context-aware responses** for employees and business users.
- Built a data ingestion and processing pipeline to extract, clean, and structure **company's knowledge files** .
- Generated **semantic vector embeddings using HuggingFace transformer models** and stored them in **PostgreSQL** for scalable retrieval.
- Implemented a **Retrieval-Augmented Generation (RAG) query system** using LangChain to fetch relevant embeddings and ground responses with company knowledge.
- Integrated **OpenAI LLM** with embedding search to provide accurate answers, reducing dependency on static FAQ systems.
- Developed and deployed a **Flask backend API** for query orchestration, embedding lookup, and real-time chatbot interactions.
- Enabled faster information access across teams, improving internal productivity and knowledge discovery.

ACHIEVEMENTS AND RESPONSIBILITIES

- **Best Contributor Award:** Recognized for exceptional contributions and consistent performance across multiple projects.
- **Adobe UX Foundation Learning Journey:** Completed the learning journey. - ([certificate](#))
- **CodeKaze Challenge:** National rank 1616 and Graduation Year rank 272 by Coding Ninjas. - ([certificate](#))
- **CodeKaze Challenge:** National rank 3106 by Coding Ninjas. - ([certificate](#))
- **Codegoda Challenge:** Global rank 505 by Agoda. - ([certificate](#))
- **Digital Marketing:** Completed "The Fundamentals of Digital Marketing" from Google Digital Garage. - ([certificate](#))
- **Nation-Wide Financial Markets Quiz Contest:** Participated in the contest. - ([certificate](#))
- **Coding Contests:** Participated in various coding contests on platforms like HackerRank, HackerEarth, CodeChef, etc.
- **Badminton:** Played at the District Level, won the Runner-up prize, and qualified for the State Level in Haridwar.