Shivang Singh

Bengaluru, India ssnfs26@gmail.com LinkedIn GitHub

Experience

Publicis Sapient June 2025 – Present
AI Engineer *Bengaluru, India*

- Architected **Bodhi Atomize** — a production multimodal GenAI platform cutting marketing asset analysis from **hours to ∼2 min per asset** (**95% reduction**) across **10,000+** assets for **Eli Lilly**; outputs 50+ structured JSON signals per asset.
- Engineered multi-stage **LLM inference pipelines** with **Gemini 2.5 Pro** and **Pydantic**-validated structured outputs; implemented token budgeting, exponential-backoff retry, and **backpressure control** to sustain production throughput under API rate limits.
- Integrated **YOLO** and **PaddleOCR** into LLM workflows, extracting 50+ typed visual components (text, characters, emotions, branding) per asset; established **LLM evaluation** with **DeepEval** (LLM-as-judge, G-Eval) to track and improve output correctness in production.
- Built **FastAPI** microservices with **Redis** (caching + task queuing) and **Celery**; deployed on **Kubernetes** with **KEDA autoscaling** to sustain **1,000+ concurrent requests** under burst traffic with low latency.

Lincode Vision Labs October 2024 – May 2025
Data Science Intern *Bengaluru, India*

- Integrated **RF-DETR** into production pipelines, achieving **1.8× faster inference** and **+7% mAP50** improvement over the YOLOv8 baseline on industrial defect detection tasks.
- Curated and preprocessed **30,000+** industrial images through targeted augmentation and annotation QA pipelines, lifting defect detection model accuracy by **10%**.

Projects

Dossier — Autonomous Agentic Job Search Intelligence System GitHub
Python, OpenAI GPT-5.4-mini/GPT-5, Claude Sonnet 4.6/Haiku 4.5, Tavily, ThreadPoolExecutor, SQLite, LaTeX

- Architected a **7-agent autonomous pipeline** (Job Discovery, Watchlist, Company Intel, Market Intel, Gap Analysis, Resume Agent, Referral Finder) that discovers, scores, researches, and generates tailored applications end-to-end — cutting per-application prep from hours to **under 2 minutes** at **$0.06/application**.
- Built **parallel LLM scoring engine** (ThreadPoolExecutor, 8 workers) across 550+ raw jobs per run with pre-LLM rule filters eliminating **65% of API calls at zero cost**; Claude Sonnet generates ATS-optimised LaTeX resumes via **3-pass self-evaluation loop** (tailor → critique → revise) with strict no-fabrication and JD keyword-mirroring constraints.
- Shipped **Referral Finder agent** with 3-tier strategy — warm LinkedIn connections → Tavily cold search with strict company + India validation → gpt-5.4-mini personalised outreach; **profile-agnostic architecture** loads all candidate data from JSON at runtime with zero code changes to switch users.

FedFV-CV — Federated Deep Learning for Biometric Authentication GitHub
PyTorch, MobileNetV2, Federated Learning

- Designed **FedFV-CV**, a federated deep learning framework for finger vein biometric authentication using **MobileNetV2**; engineered custom **FedWPR** aggregation on **122,600 images** across 5 clients, achieving **1.21% EER** — outperforming FedAvg benchmarks. (B.Tech Thesis, IIIT SriCity)

slackAgent — AI-Powered Slack Bot GitHub
FastAPI, LlamaIndex, ChromaDB, OpenAI, n8n

- Built a scalable **FastAPI** backend with **LlamaIndex** + **ChromaDB** semantic search over 20+ documents; cut query response time by **40%** and served **50+ daily queries** via Slack API with end-to-end workflow automation using **n8n**.

RAG-QA Chatbot on AWS GitHub
LangChain, FAISS, AWS Bedrock, LLAMA 3.1-70B, Docker, GitHub Actions

- Built a retrieval-augmented QA system using **LangChain**, **FAISS**, and **AWS Bedrock** (LLAMA 3.1-70B); deployed to AWS ECR + App Runner via **Docker** with full CI/CD through **GitHub Actions**.

Technical Skills

Programming & ML: Python, PyTorch, scikit-learn, pandas, NumPy, SQL
LLM & GenAI: Gemini 2.5 Pro, OpenAI, LangChain, LangGraph, LlamaIndex, RAG, Prompt Engineering, Pydantic
Computer Vision: YOLO, RF-DETR, PaddleOCR, OpenCV
MLOps & Backend: FastAPI, Docker, Kubernetes, KEDA, Redis, Celery, DeepEval, mlflow
Cloud & Infrastructure: GCP, AWS (Bedrock, ECR, App Runner), Azure, GitHub Actions, ChromaDB, FAISS

Education

Indian Institute of Information Technology, SriCity 2021 – 2025
Bachelor of Technology in Computer Science and Engineering *CGPA: 8.09 / 10*

Involvement

Computer Vision Lead — Matrix Club, IIIT SriCity *2023 – 2025*
State-Level Table Tennis Player IIIT SriCity Table Tennis Club