Shivansh Pareek

Generative AI — Backend Development— Cloud Engineering

☎ +91 8448779855 ✉ shivanshp99@gmail.com Delhi, India
linkedin.com/in/shivanshpareek github.com/shivzz-creator Portfolio Website

SUMMARY

AI Engineer with strong experience in Generative AI, backend engineering, and cloud-native development. Skilled in building end-to-end LLM systems, vector search pipelines, scalable APIs, and production-grade automation workflows. Delivered commercialization-ready AI solutions across Samsung and IBM, including template generation engines, text–video similarity pipelines, and WatsonX-driven asset productivity tools. Hands-on with Python, Node.js, AWS, IBM Cloud, embeddings, RAG, HuggingFace, and modern agentic frameworks; experienced in shipping enterprise AI features from research to deployment.

EXPERIENCE

IBM Bangalore, India
AI Engineer – WatsonX CE Nov 2025 – Present

- Supported the Annotate Asset Productivity initiative by improving asset insights using WatsonX Runtime, MCP, and Code Engine.
- Completed Watson Orchestrate L2–L4, Agentic AI, and RAG trainings; contributed to early-stage AI automation workflows.

Samsung Research Institute - Delhi Delhi, India
AI Software Engineer Jan 2024 – Oct 2025

- Designed, implemented, and optimized backend APIs in Node.js, including a context-aware search API that streamlined retrieval of images, videos, and media assets across enterprise platforms, reducing manual lookup time by 35%.
- Configured AWS SQS and developed a robust listener service to track real-time content updates. Enhanced the uploader service to process and enrich video/image metadata, enabling downstream analytics and automated workflows.
- Built and deployed a corporate template generator using LangChain and LLMs. Delivered 4+ custom corporate templates in under 16 seconds by integrating smart cropping, scene detection, and masking—accelerating internal content creation by 40%.
- Architected a modular backend in Flask/FastAPI, containerized with Docker, with dedicated services for database access, file storage in S3, LLM interaction, and prompt generation. Integrated DB session pooling and multithreading, improving throughput by 25%.
- Designed and deployed a text-to-video similarity pipeline using FAISS and PostgreSQL; processed over 1,000+ videos with average <4s response time by leveraging frame selection and semantic embeddings.
- Collaborated with R&D teams on Content Transformation using SDXL (Stable Diffusion), segmentation, and comparative analysis; contributed to AWS Lambda workflows and LLM fine-tuning pipelines.
- Mentored and supervised interns on tasks such as smart cropping, vector search, and backend service design, improving code quality and ensuring knowledge transfer across teams.

Pepcoding Ltd. Noida (Remote)
Backend Developer Intern Jan 2022 – Jun 2022

- Built REST APIs for the NADOS learning portal serving 10,000+ weekly users with 99.9% uptime.
- Automated workflows using Puppeteer, saving 20+ developer hours monthly; optimized DB queries reducing server load by 18%.
- Mentored 100+ students weekly in DSA and interview preparation with a 4.8/5 feedback score.

SKILLS

- **Languages:** Python, Java, JavaScript, SQL
- **AI/LLM:** RAG, Embeddings, LangChain, Vector Search, SDXL, Agentic AI
- **Backend:** Flask, FastAPI, Node.js, Express.js, REST APIs
- **Cloud & DevOps:** AWS (EC2, S3, RDS, SQS, Lambda, Bedrock), IBM Cloud, Code Engine, Docker, Git, Jenkins
- **Databases:** PostgreSQL (pgvector), MySQL, MongoDB
- **Others:** MCP, Architecture Design, Mentoring, Problem Solving
- **WatsonX Ecosystem:** WatsonX Runtime, WatsonX.ai, Watson Orchestrate

PROJECTS

IPL Stats Analyzer & Chatbot *Node.js, Python, FAISS, Streamlit, LangChain*
Project Link

- Scraped and structured IPL data using Node.js & Cheerio.
- Developed a LangChain-based RAG chatbot for contextual cricket analytics.
- Implemented FAISS-based semantic vector retrieval and embedding pipeline.
- Built Streamlit dashboards for match insights and player-level analytics.

ACHIEVEMENTS

- **Open Lab Innovation:** Commercialized a Generative AI-based product at Samsung.
- Ranked Top 30% in Samsung's SWCProf Certification.
- **GenAI Master (Level 2):** Scored 98%; appointed as AI Moderator internally.
- Solved 500+ DSA problems across LeetCode, GFG, and InterviewBit.
- Completed 100+ hours of AI upskilling under IBM Super Learner Program.

EDUCATION

The LNM Institute of Information Technology (LNMIIT) Jaipur, India
B.Tech in Computer and Communication Engineering Jul 2019 – Jul 2023
CGPA: 7.63

N.K. Bagrodia Public School Delhi, India
Class 12: 94.3% (CBSE)
Class 10: CGPA 10.0