SHIVANSH MAHAJAN

Patna, India | +91 9798940184 | shivansh.m2003@gmail.com
LinkedIn | GitHub | LeetCode

PROFESSIONAL SUMMARY

Innovative AI/ML Engineer with expertise in Generative AI, LLM applications, and RAG pipelines. Currently pursuing BTech in Computer Science at JUET with proven track record of building production-grade AI systems. Proficient in LangChain, LangGraph, FastAPI, and cloud technologies. Demonstrated ability to architect multi-agent workflows, optimize retrieval accuracy, and deliver impact in real-world AI applications.

WORK EXPERIENCE

AI Intern at Zeron

Focus: Multi-agent workflows, RAG pipeline optimization, autonomous AI agents

• Architected 2 multi-agent workflows and 2 RAG pipelines (corrective & agent-based) that lifted retrieval accuracy from 71% → 93% using hybrid search, ranking, and MMR

• Built custom MCP servers on Render and 3 LangFlow AI workflows; launched ZIN AI, an autonomous chatbot serving as single AI assistant across all Zeron client products

• Presented bi-weekly demos to CTO and senior engineers—translated technical trade-offs into ROI metrics, securing production launch approval

• Collaborated with cross-functional teams resulting in 30% increase in operational efficiency across projects

• Spearheaded AI model research and optimization, improving model accuracy by 25% through iterative testing

• Built efficient AI-powered SCP mapping with frameworks like NIST and SEBI

AI Software Development Intern at Stremly

Duration: April 14 – June 14, 2025 | **Focus:** Web automation, AI agents, production systems

• Co-engineered production-grade agent-based web-automation platform with 4 interns and CTO within 8-week sprint

• Refactored single-threaded prototype into 4 autonomous agents (Crawler, Extractor, Validator, Critique) leveraging Graph-RAG for DOM element pinpointing

• Designed and implemented components for AI agent workflows with LLMs, vector stores, and prompt engineering

• Integrated automation flows involving browser and system-level interactions using Playwright

• Tech Stack: LangGraph, Playwright, LangChain, Pinecone, FastAPI, Pydantic

PROJECTS

1. Voice-Activated Portfolio Assistant (Interview Prep)

Real-time voice-activated AI interview assistant using LiveKit's agent framework for technical interview preparation

• Speech-to-Text (STT): Deepgram Nova-3 with real-time interim results

• LLM: OpenAI GPT-4.1-mini for context-aware responses

• Text-to-Speech (TTS): OpenAI TTS with natural voice output

• VAD: Silero VAD with optimized thresholds (300ms min speech, 500ms silence, 0.5 activation)

• RAG Pipeline: Pinecone vector DB + OpenAI embeddings for semantic search over personal knowledge base

• Sub-2 second latency, mock interview capabilities, hands-free operation for commuting scenarios

• Production-ready Docker deployment with comprehensive logging and error handling

2. Fraud Detection System (Deep Learning)

ML-based fraud detection API predicting fraudulent financial transactions using deep neural networks

• Deep Neural Network: 4 hidden layers with batch normalization and dropout regularization
• Input: 12 engineered features (transaction amount, frequency, distance, credit score, etc.)
• SMOTE-based class imbalance handling; 5-fold Stratified K-Fold CV
• Metrics: ~8-11ms prediction latency, ~100 transactions/second throughput
• REST API: Flask + CORS support with comprehensive error handling
• Deployment: Gunicorn production WSGI server, Docker containerization ready

3. LinkedIn Blog Agent

AI-powered assistant transforming multi-format content into engaging LinkedIn posts using vision and language models

• Multi-format support: PDFs, images, code files (20+ languages), presentations, text
• Vision Analysis: Google Gemini Flash 1.5 for visual content understanding
• Code Analysis: Anthropic Claude for technical content extraction
• LLM: Anthropic Claude Opus for high-quality blog generation
• Advanced presentation processing: PowerPoint/PDF extraction with speaker notes
• Human-in-the-loop generation: LangGraph-based interactive refinement with regeneration options
• LinkedIn optimization: Viral hooks, hashtags, CTAs, emoji usage, posting best practices

4. Interactive Storytelling API (FastAPI Backend)

Dynamic AI-powered story creation backend with advanced character development and professional exports

• Multi-dimensional character profiles: Up to 10 characters with relationships and AI-generated backstories
• 6 theme-based story creation: Fantasy, Mystery, Adventure, Sci-Fi, Horror, Romance
• Dual-mode generation: User-guided choices or AI auto-continuation
• Real-time paragraph editing with natural language instructions
• Professional exports: High-quality PDF generation, multi-language audio narration (10+ languages)
• 100% test coverage (18/18 tests passing), 8ms average response time
• Tech: FastAPI, Pydantic, LangChain, Groq API, FPDF, gTTS

5. Ultimate Summarization API (Multi-format)

Comprehensive backend processing and summarizing legal documents, general docs, resumes, audio, video, and websites

• Legal Document Module: Document type detection, map-reduce summarization, Tavily legal context enrichment
• General Documents: Dynamic strategy (short vs long), section importance scoring, configurable summaries
• Resume Analysis: Structured data extraction, ATS compatibility analysis, job description comparison
• Audio Processing: AssemblyAI transcription, recursive chunking, map-reduce summarization
• Video Processing: YouTube and uploaded video analysis with multi-model integration
• Website Processing: Async crawling via crawl4ai, customizable summary lengths, Groq Gemma2 model
• Dependency injection with singleton pattern, comprehensive error handling, background task cleanup

6. Data Analyst AI Assistant (Streamlit + LangChain)

Intelligent data analysis platform with interactive visualizations and natural language Q&A; interface

• LangChain agents with specialized DataFrame and visualization tools

- Multi-format support: CSV, Excel, PDF, images
- Auto-generates 20+ Plotly charts (numerical, categorical, mixed) on demand
- Conversation-buffer memory: Natural language Q&A; up to 20 human messages
- Dual agent implementations: LangChain and LangGraph architectures
- Statistical analysis, time series analysis, relationship discovery, anomaly detection
- Tech: Python 3.12, LangChain, LangGraph, Streamlit, OpenAI GPT-4o, Gemini-Flash-2.0, Plotly

HACKATHON EXPERIENCES

1. InnovateX Delhi

Organization: Delhi Technological University (DTU)
Achievement: Participated & Showcased
Project: Fraud Detection System

2. Hackout Hackathon

Organization: Dhirubhai Ambani University (DA-IICT)
Achievement: Finalist - Top 10 out of 450 teams
Project: Agriculture-based AI Project

3. Hack The Mountains by MLH

Organization: Major League Hacking (MLH)
Achievement: Selected & Participated
Project: AI Meeting Platform

4. CraveFeed - Phase 2 Ride Hacks

Organization: JIIT Noida
Achievement: ■ 1st Runner-Up
Project: CraveFeed

EDUCATION

Bachelor of Technology (B.Tech)	Jaypee University of Engineering Technology (JUET)	Guna, Madhya Pradesh	Sep 2022 – May 2026	CGPA: 7.1/10
12th Grade	Kautilya Senior Secondary School	Kota	2019-2021	78%
10th Grade	Don Bosco Academy	Patna	2019	90%

SKILLS

Generative AI & LLM Frameworks: LangChain, LangGraph, Agno, Crew AI, Hugging Face, Google ADK, LlamaIndex, Guardrails, MCP Protocol, A2A Protocol

Cloud & Databases: AWS (EC2, S3, SageMaker, Bedrock, Lambda), Supabase, MongoDB, Redis, Pinecone, Chroma DB, Snowflake

AI/ML Technologies: Generative AI, NLP, Deep Learning, LLM Fine Tuning, RAG Techniques, Machine Learning, Data Analysis, AI Agents, LLM Monitoring, CI/CD

Python Libraries & Frameworks: Pandas, Matplotlib, Seaborn, Scikit-learn, TensorFlow, PyTorch, FastAPI, NLTK, Spacy, Beautiful Soup, Selenium, FastMCP, Graphiti, Crawl4AI, Plotly, Streamlit, Pydantic

Tools & Software: Tableau, GitHub, Docker, MLFlow, GitHub Actions, LangFlow, LangSmith, LiveKit, Streamlit

Competitive Programming: 300+ LeetCode problems solved, competitive programming experience, algorithm optimization