Shivam Agarwal

+91-6398105401 shivamagarwal2211@gmail.com linkedin.com/in/shivam-agarwal github.com/shivam0897-i

Professional Summary

Results-driven AI/ML Engineer with proven expertise in developing production-ready intelligent systems that deliver measurable business impact. Specialized in building LLM-powered applications, computer vision models, and multimodal AI platforms using Python, TensorFlow, and Hugging Face. Demonstrated success in achieving 92%+ precision rates and sub-2 second API latencies through optimized ML pipelines. Experienced in deploying scalable AI solutions on GCP and Firebase, with strong focus on automation, efficiency, and real-world problem solving.

Technical Skills

Programming: Python, JavaScript, TypeScript, SQL, Git/GitHub, REST APIs
Machine Learning: TensorFlow, Keras, PyTorch, Scikit-learn, Transfer Learning, Model Fine-tuning, Hyperparameter Optimization
Deep Learning & CV: ResNet50, CNNs, Computer Vision, OpenCV, Image Classification, Object Detection, Data Augmentation
Generative AI: Google Gemini, LangChain, LiteLLM, Stable Diffusion XL, Hugging Face Transformers, Prompt Engineering, RAG
Data Engineering: NumPy, Pandas, FAISS Vector Databases, PyPDF2, Semantic Search, Data Preprocessing, Feature Engineering
Cloud & MLOps: Google Cloud Platform (GCP), Firebase, Vercel, AWS, Docker, Model Deployment, API Development
Web Technologies: React.js, FastAPI, Streamlit, Asynchronous Processing

Education

Meerut Institute of Engineering and Technology (MIET) **Meerut, Uttar Pradesh**
Bachelor of Technology in Computer Science & Engineering - AI & ML Specialization *2023 – 2027*

Professional Experience

AI/ML Engineer Intern **August 2025 – Present**
Point9 *Remote*

- Architected and deployed 3+ custom AI agents using LiteLLM, automating enterprise workflows for cheque processing, KYC verification, and legal document analysis, reducing manual processing time by 70%
- Engineered intelligent cheque processing agent with OCR and validation capabilities, achieving 95%+ accuracy in data extraction and enabling real-time fraud detection across 10,000+ daily transactions
- Developed multilingual legal document summarizer supporting 5+ languages, processing 500+ page documents in under 30 seconds while maintaining context accuracy and compliance with regulatory standards
- Designed scalable end-to-end AI pipelines with LiteLLM orchestration, optimizing API response times by 40% and reducing infrastructure costs by 25% through efficient resource allocation

Key Projects

ThinkPDF – AI-Powered Document Intelligence Platform | *Google Gemini, FAISS, Python, PyPDF2*

- Built production-grade semantic search platform processing 1,000+ multi-document queries daily with Google Gemini LLM and FAISS vector databases, achieving 92%+ retrieval precision and sub-500ms query response times
- Implemented advanced chunking algorithm optimizing context window utilization, enabling accurate extraction from 100+ page documents with 85% reduction in hallucination rates compared to baseline models
- Developed interactive conversational interface supporting multi-turn dialogues and automated summarization, improving user productivity by 60% through intelligent document navigation and topic extraction

MultiModal AI Content Generation Suite | *Gemini 2.5 Flash, Stable Diffusion XL, Hugging Face*

- Engineered enterprise-ready multimodal platform integrating text generation, image synthesis, and audio processing, serving 500+ API requests daily with 99.8% uptime and sub-2 second latency
- Optimized image generation pipeline using asynchronous processing and PIL streaming, reducing memory footprint by 45% and enabling concurrent processing of 50+ requests without performance degradation
- Integrated 6+ AI APIs (Gemini 2.5, Stable Diffusion XL, Google Translate) into unified workflow, cutting development time for new features by 40% through modular architecture and comprehensive error handling

AI-Powered Waste Classification System | *ResNet50, TensorFlow, Keras, Streamlit*

- Fine-tuned ResNet50 architecture on 2,500+ labeled images achieving 47.5% accuracy on TrashNet dataset through transfer learning, data augmentation, and hyperparameter optimization, outperforming baseline by 12%
- Deployed real-time classification system with Streamlit interface processing images in ¡100ms on edge devices, enabling instant recycling recommendations and reducing sorting errors by 35%
- Implemented model compression techniques reducing model size by 60% while maintaining accuracy, making solution viable for deployment on resource-constrained IoT devices and mobile applications

Certifications & Awards

Oracle Cloud Infrastructure 2025 Certified Generative AI Professional | Google Cloud Generative AI Virtual Internship | IBM AI Fundamentals | Gemini Certified University Student | Python Full Stack Development