KUMAR SHANU

Generative AI Engineer | LLM Developer | AI Engineer

📧 Kumarshanu2110@gmail.com | 📱 9599120357 | 📍 New Delhi | Linkedin

PROFESSIONAL SUMMARY

Generative AI Engineer with 3+ years of hands-on experience in developing, fine-tuning, and deploying large language models and generative AI systems. Specialized in creating and constructing new models for text, image, and video generation by adjusting core models to specific use cases. Expert in GitHub Copilot implementation, Gemini AI integration, and building scalable AI-powered developer tools that enhance productivity and streamline development workflows.

Core Expertise: LLM Fine-tuning, Advanced Prompt Engineering, AI Agent Development, Model Optimization, Production AI Systems, RAG Implementation, Multi-modal AI

TECHNICAL SKILLS

Generative AI & Large Language Models: OpenAI GPT-4/4o, Google Gemini Pro/Ultra, Claude 3.5 Sonnet, Llama 2/3, Code Llama, LangChain, LlamaIndex, Hugging Face Transformers, Ollama, AutoGen, Fine-tuning, RLHF, RAG, Few-shot learning, Zero-shot learning, Chain-of-thought, Constitutional AI

Vector Databases & Storage: Pinecone, ChromaDB, Weaviate, FAISS

AI-Powered Development Tools: GitHub Copilot, Gemini Code Assist, OpenAI Codex, Google AI Studio, AI Agents, Task Automation, Code Generation

Programming & Development: Python 3.11, JavaScript, SQL, Bash, PyTorch, TensorFlow, scikit-learn, NumPy, Pandas, Django, Flask,

Cloud & Infrastructure: Azure OpenAI, Google Vertex AI, AWS Bedrock, Azure AI, Google Cloud AI Platform, Hugging Face Hub, Docker, Kubernetes, CI/CD, Git, GitHub Actions

Databases & Monitoring: PostgreSQL, MongoDB, Redis, Vector Databases, MLflow, Weights & Biases, TensorBoard

PROFESSIONAL EXPERIENCE

Senior Generative AI Engineer | Infosys Limited

Chandigarh | October 2023 – March 2025

GitHub Copilot & AI Leadership:

- Built Copilot-powered agents for brainstorming, planning, building, testing, and running code in natural language, serving 50+ developers with 95% satisfaction rate
- Developed advanced prompt engineering techniques for GitHub Copilot, improving code suggestion accuracy by 60% and reducing debugging time by 35%
- Created comprehensive automated testing frameworks using AI agents, achieving 85%+ code coverage and 50% faster test execution

LLM Fine-tuning & Model Optimization:

- Investigated, created, and constructed new generative models for text, images, and code by fine-tuning core models for specific use cases
- Led the creation of pipelines for cleaning and preparing data to facilitate model training for 5M+ records
- Implemented model quantization and optimization techniques, reducing inference time by 50-70%
- Scaled AI models to handle 10K+ daily active users with 99.9% uptime

Gemini AI Integration & Development:

- Built production applications using Gemini Code Assist and Google AI Studio
- Developed applications combining text, code, and visual inputs using Gemini Pro
- Created efficient API integration patterns for Gemini models, reducing response time by 45%

AI/ML Systems Engineer | Infosys Limited

Chandigarh | May 2022 – September 2023

Generative AI Pipeline Development:

- Designed and optimized generative models for tasks such as natural language generation and data augmentation
- Reviewed data, refined algorithms, and trained models to improve accuracy across multiple AI projects
- Attended meetings with cross-functional teams to align on project goals and troubleshoot issues
- Created intelligent feature engineering algorithms, reducing model training time by 60%

AI/ML Engineering Trainee | Infosys Limited

Chandigarh | February 2022 – April 2022

Intensive AI Specialization:

- Completed accelerated Python and Machine Learning certification with top performance
- Demonstrated expertise in Python 3.11, NumPy, Pandas, scikit-learn, and advanced ML algorithms
- Established strong fundamentals in generative AI, model development, and deployment

KEY PROJECTS

AI-Powered Code Generation Platform

Technologies: GitHub Copilot, OpenAI Codex, LangChain, FastAPI

- Developed intelligent code generation system that adjusts core models to specific use cases
- Created sophisticated prompt templates reducing development time by 30-40%
- Built seamless integration with existing development workflows and IDEs
- Implemented AI-powered code review and testing automation

Multi-Modal AI Assistant (Gemini Integration)

Technologies: Google Gemini Pro, AI Studio, Vector DB, React

- Built conversational AI system handling 1K+ daily interactions
- Developed system for analyzing and explaining complex codebases
- Created interface for generating code from natural language descriptions
- Achieved 95% user satisfaction with sub-2-second response times

Enterprise AI Agent Framework

Technologies: LangChain, Azure OpenAI, Docker, Kubernetes

- Developed agents for streamlining development workflows by automating tasks such as bug fixes and feature implementations
- Created system for developers to delegate complex tasks to AI agents
- Implemented autonomous agents working in background while developers focus on core tasks
- Designed system handling 100+ concurrent AI agent operations

CERTIFICATIONS

- Infosys Certified Generative AI Professional – Advanced Level
- Infosys Certified Applied Generative AI Professional
- AI-900 - Microsoft Azure AI Fundamentals

EDUCATION

Bachelor of Technology (B.Tech)
Bhagwan Mahaveer College of Engineering and Management, Sonipat
2021 | CGPA: 7.05/10.0

Senior Secondary (Class XII)
YKJM School
2017 | 60.4%

Secondary (Class X)
Gyan Mandir Public School
2014 | CGPA: 7.0/10.0

ACHIEVEMENTS

Innovation & Leadership:

- AI Transformation Leader: Spearheaded enterprise-wide AI adoption, training 50+ developers
- Performance Excellence: Achieved 40% productivity improvement through AI tool implementation
- Research Contributions: Published 3 internal research papers on prompt optimization and model fine-tuning
- Open Source: Active contributor to LangChain and Hugging Face communities

Business Impact:

- Cost Reduction: Implemented AI optimization reducing cloud infrastructure costs by 40%
- Developer Productivity: Boosted developer productivity through strategic GitHub Copilot implementation