VINAYAK GALANDE

AI Engineer / Software Engineer

CONTACT

7559237537

vinayakpgalande90@gmail.com

LinkedIn

GitHub

TECHNICAL SKILLS

- NLP & Generative AI: LLMs, RAG, Hugging Face, LangChain | Embeddings, Text Extraction, NLTK
- Machine Learning & Deep Learning: scikit-learn, TensorFlow, Keras | CNN, ANN, RNN, LSTM
- Programming & Databases: Python, SQL, MySQL
- Vector Databases & Retrieval : FAISS, Semantic Search
- Data Analysis & Visualization: EDA, Data Cleaning, Feature Engineering | NumPy, Pandas, Matplotlib
- Backend & Deployment: REST APIs, FastAPI

EDUCATION

- B.tech | 2020 - 2024
  SPPU **| 8.27 CGPA**

SOFT SKILLS

- Problem-solving
- Adaptable
- Team Collaboration
- Attention to Detail
- Clear Communication

ACHIEVEMENTS & CERTIFICATIONS

- Coursera: Biology Meets Programming: Bioinformatics for Beginners
- Online Workshop: Genome Informatics: Decode life
- Hacker Rank: 4-Star in Python (ongoing)
- HackerRank: 2-Star in SQL (ongoing)

PROFILE

Hands-on AI Engineer with 1+ year of industry experience building LLM-based agents and RAG pipelines, and deploying ML systems via FastAPI. Experienced in NLP, embeddings, and vector databases, with a focus on shipping AI solutions that drive automation and data-driven decision-making.

EXPERIENCE

iSPEEDBiz Pvt. Ltd | May 2025 - Present

Role : Software Engineer

- Developed and integrated RESTful APIs using Flask/FastAPI to support AI-powered chatbot and LLM-based modules, enabling seamless communication between backend services and a React-based frontend.
- Implemented backend logic for prompt handling, response orchestration, and API-level integration of AI components.
- Worked closely with senior engineers on debugging, performance optimization, and modular deployment, ensuring stable and scalable backend services.

CodeSpyder Technologies Pvt. Ltd | Aug 2024 - Feb 2025

Role : Data Scientist Intern

- Designed and implemented end-to-end Machine Learning and Deep Learning pipelines using TensorFlow and Keras, covering data preprocessing, EDA, feature engineering, and hyperparameter tuning to improve model performance and reliability.
- Built NLP and Generative AI components, including RAG pipelines, text embeddings, and vector-based retrieval, to enable intelligent document analysis and resummarization use cases.
- Integrated FAISS-based vector search with LLM workflows to enhance semantic retrieval accuracy across large unstructured text datasets.

PROJECTS

- **Agentic AI**
  - Designed and implemented an agent-driven AI backend using FastAPI, enabling LLM-based modules to handle multi-step task execution, response orchestration, and contextual decision-making.
  - Architected a multi-agent workflow layer with API-level abstractions for prompt management, agent coordination, and tool invocation, enabling modular LLM-based agent interactions.
  - Implemented document ingestion and semantic retrieval by indexing company documents into a Milvus vector database, supporting agent context grounding and scalable backend workflows.
- **LLM-based RAG System for Research Summarization**
  - Built an end-to-end RAG system using LangChain, FAISS, and Hugging Face embeddings to enable semantic search and context-aware summarization over research papers.
  - Optimized text chunking and vector retrieval to improve relevance and reduce hallucinations in LLM-generated outputs.
  - Reduced reliance on paid APIs by integrating open-source embeddings, improving cost efficiency while maintaining accuracy; currently extending toward multi-document analysis.
- **Semantic Text Analysis using LLMs, ML & Deep Learning**
  - Developed a semantic sentiment analysis pipeline using BERT embeddings, benchmarking ML models (Logistic Regression, SVM, Gradient Boosting) against a custom DL model.
  - Achieved ~78% accuracy through optimized preprocessing, batching, subsampling, and hyperparameter tuning.
  - Conducted controlled experiments to analyze embedding effectiveness and model trade-offs, ensuring reproducible evaluation.