

Abhivanth R

abhocoder39@gmail.com | github.com/Abhivanth-08 | linkedin.com/in/abhivanth-r | abhivanth-r.vercel.app

Professional Summary

ML Engineer specializing in Generative AI, Large Language Models, and RAG systems. Expertise in building production-ready AI applications using LangChain, transformers, and vector databases. Experienced in developing intelligent agents, NLP pipelines, and deploying scalable ML models. Published researcher with hands-on experience in LLM inference, prompt engineering, and AI-powered automation. Strong foundation in deep learning, computer vision, and end-to-end ML deployment.

Technical Skills

GenAI & LLMs: LangChain, Transformers, GPT-2, BERT, Prompt Engineering, RAG Systems

ML/DL Frameworks: PyTorch, Scikit-learn, TensorFlow, Hugging Face

NLP: spaCy, NLTK, Text Processing, Entity Recognition, Semantic Search

Vector Databases: LanceDB, Embeddings, Similarity Search

Languages: Python, SQL, R, TypeScript

MLOps: FastAPI, Flask, Docker, Git, Model Serving, API Development

Document Processing: PyMuPDF, Docling, PDF Parsing, OCR

Cloud & Tools: Azure AI, HuggingFace, Vercel, Supabase

Experience

ML Engineer Intern

January 2025 – February 2025

Nitroware Technologies, Coimbatore

- Developed machine learning models for predictive breach event detection with Flask-based API deployment and real-time inference
- Implemented end-to-end ML workflows including model serving infrastructure, performance monitoring, and optimization strategies

Projects

LLM Internals Explorer

[Demo](#)

- Built transformer architecture visualization platform with real GPT-2 (124M) and BERT inference featuring Attention Head Personality Profiler analyzing 144 attention heads using t-SNE clustering on high-dimensional attention patterns
- Implemented semantic vector arithmetic (King - Man + Woman = Queen), real-time BPE tokenization analysis, and prompt injection detection system using PyTorch and Transformers library

College Chatbot (RAG System)

[Demo](#)

- Engineered production-grade RAG chatbot using LangChain for intent classification and LanceDB vector store for semantic retrieval automating admissions, events, food ordering, and exam preparation
- Built document processing pipeline with Docling for multi-format parsing and context extraction, handling multi-turn dialogues with conversation state management

Info Redaction Agent

[Demo](#)

- Developed AI-powered PII detection system using NLP pipeline (spaCy, LangChain) for entity recognition processing PDFs and structured data with context-aware redaction and GDPR/CCPA compliance
- Implemented intelligent pattern matching engine with document parsing (PyMuPDF, Docling) and automated privacy validation achieving high precision on diverse PII types

Pull Request Review Agent

[Demo](#)

- Built AI code review system using LangChain reasoning engine for automated PR analysis, identifying security vulnerabilities and code quality issues with actionable feedback generation
- Architected LLM-powered code understanding pipeline with AST analysis, multi-language support, and GitHub API integration for seamless developer workflow automation

AI Foley Studio

[Demo](#)

- Developed multimodal AI system combining computer vision (YOLO, MediaPipe) and LangChain decision engine for intelligent sound effect selection and frame-accurate audio synchronization
- Built end-to-end ML pipeline for video analysis, action detection, and context-aware audio generation with real-time processing capabilities

Publications

Secure and Ethical AI-Driven Approaches to Data Privacy in Sustainable Agriculture

September 2025

CRC Press Book Chapter (Accepted) – *Blockchain and AI for Security and Privacy in Smart Agriculture*

Education

K P R Institute of Engineering and Technology

2023 – 2027

B.E. Computer Science and Engineering (AIML Specialization)

CGPA: 8.8/10