NAGAVARDHAN BATTU

Maineville, OH |+1 (989) 854-9626 | nagavardhan768@gmail.com | LinkedIn | GitHub | Portfolio

PROFESSIONAL SUMMARY

Results-driven Generative AI Engineer with 3+ years of experience designing, developing, and deploying LLM-powered, RAG-based, and multimodal AI solutions across financial and enterprise domains. Skilled in Python, PyTorch, TensorFlow, LangChain, LangGraph, Hugging Face, OpenAI, and Azure AI Services. Experienced in building multi-agent LLM workflows, fine-tuning domain-specific models, integrating vector databases (Milvus, FAISS, ChromaDB), and implementing OCR and multilingual processing to automate workflows and enhance enterprise decision-making. Proven ability to deliver scalable AI platforms, monitor model performance with dashboards, and translate advanced AI research into production-grade systems with measurable impact.

TECHNICAL SKILLS

Programming & Frameworks: Python, Java, SQL, FastAPI, Flask, Streamlit, LangServe, Gradio, PyTest

Generative AI & LLMs: GPT-4, Claude, Gemini, Llama 3, Mistral, RAG Pipelines, LangChain, LangGraph, Hugging Face Transformers, Prompt Engineering, Function Calling, Tool Integration, AI Agents, Agentic Workflows, Tavily, Groq LLaMA-3.1

Vector Databases & Retrieval: ChromaDB, Milvus, FAISS, Pinecone, Weaviate, Neo4j, Redis Vector, Elasticsearch

Machine Learning & Data Science: Supervised & Unsupervised Learning, Predictive Modeling, Feature Engineering, Scikit-learn, NumPy, Pandas, XGBoost, LightGBM, MLflow, Model Evaluation, Experiment Tracking

Deep Learning & NLP: Transformers, Encoder-Decoder Models, BERT, T5, LlamaIndex, Text Generation, Summarization, Fine-Tuning (LoRA, QLoRA, PEFT), TensorFlow, PyTorch

Multimodal & Synthetic AI: CLIP, BLIP-2, DALL·E, Whisper, Stable Diffusion, Vision Transformers (ViT), Speech-to-Text, Image-to-Text, OCR (Tesseract), Multimodal RAG

Cloud & MLOps: Azure AI Services, Azure Machine Learning, AWS SageMaker, GCP Vertex AI, Docker, Kubernetes, GitHub Actions, CI/CD Pipelines, Model Deployment & Monitoring, Weights & Biases

Visualization & Analytics: Power BI, Tableau, Looker Studio, Plotly, Grafana, Interactive Dashboards

Development Tools & Collaboration: Jupyter Notebook, VS Code, PyCharm, Git, GitHub, Notion, Confluence, Agile Workflow, API Testing (Postman)

PROFESSIONAL EXPERIENCE

Generative AI Engineer **Sep 2024 - Present**
OneMain Financial | USA

- Spearheaded the creation of a document intelligence platform using Python, Azure AI Vision, and LangGraph, transforming a prototype into an enterprise-grade solution that accelerated processing of PDFs, JSON, and scanned files by 3×.
- Developed RAG pipelines with LangGraph, Hugging Face embeddings, and Milvus, enhancing search precision by 45% and recall by 42%, enabling faster and more accurate retrieval of critical financial data.
- Designed multi-agent LLM workflows combining retrieval, reasoning, summarization, and citation verification with PyTorch and LangChain, cutting manual document review by 60% while delivering explainable and trustworthy AI outputs.
- Expanded document coverage by integrating OCR and multilingual text extraction using Azure AI Vision and Tesseract, supporting over 10 languages and boosting ingestion speed by 35% across enterprise datasets.
- Implemented active learning mechanisms to capture user feedback and dynamically adjust embeddings with Python and LangGraph, resulting in a 25% improvement in contextual relevance for complex financial queries.
- Applied software engineering best practices such as modular design, OOP, unit testing, and CI/CD pipelines, improving platform scalability and reducing deployment issues by 30% for production AI services.
- Created interactive dashboards with Streamlit and Power BI to monitor query patterns, latency, and accuracy, providing actionable insights that increased adoption and operational efficiency by 30%.
- Fine-tuned domain-specific LLMs for financial and healthcare datasets using PyTorch and TensorFlow, enhancing model performance by 15–20% and enabling more reliable predictions for enterprise applications.

Generative AI Intern **May 2024 - Aug 2024**
OneMain Financial | USA

- Led the creation of a Generative AI Research Assistant using LangGraph, orchestrating retrieval, reasoning, summarization, and citation verification workflows, reducing research turnaround time by 70%.
- Leveraged Tavily Search API and Groq LLaMA-3.1 to deliver low-latency contextual search and text synthesis, producing explainable outputs with citations and increasing research accuracy by 40%.
- Designed a Streamlit-based interactive dashboard enabling users to query topics, track multi-agent workflow progress, and export structured AI-generated reports, improving adoption among business teams by 50%.
- Constructed modular LLM components with version-controlled pipelines and environment configurations, allowing seamless replication across environments and cutting deployment setup time by 30%.
- Optimized LLM reasoning and memory persistence using prompt chaining and context management, enhancing multi-step query continuity and boosting factual consistency across research sessions.

- Authored detailed documentation, architecture diagrams, and performance metrics in Python notebooks and Confluence, ensuring smooth knowledge transfer and supporting transition to production-ready deployment.

LLM Engineer **Jan 2021 - Jun 2023**
Cognizant | India

- Led development of LLM-powered analytics pipelines using Python and Hugging Face Transformers, automating KPI extraction and cutting manual reporting by 40%, enabling stakeholders to access real-time insights.
- Built robust data preprocessing and validation workflows with Python and SQL, ensuring 99% accuracy across datasets used for training and fine-tuning enterprise LLMs powering predictive models.
- Leveraged NLP techniques like sentiment analysis, topic modeling, and entity recognition using Scikit-learn and spaCy, generating actionable insights that enhanced model performance and informed business strategies.
- Collaborated with Data Science teams to curate and feature-engineer domain-specific datasets, boosting LLM relevance and improving model accuracy by 25% across enterprise applications.
- Engineered integration of Java APIs with SQL and MongoDB to enable high-performance LLM inference workflows, reducing query latency by 30% and improving end-to-end system responsiveness.
- Designed and deployed ETL pipelines for structured and unstructured data with Azure Data Factory, streamlining ingestion and preprocessing of multimodal datasets for LLM training and evaluation.
- Implemented RAG pipelines using vector databases such as Milvus and FAISS, improving enterprise knowledge retrieval and decreasing manual lookup effort by 50%.
- Developed Power BI dashboards to track LLM performance, adoption, and output accuracy, enabling data-driven decisions and accelerating business insights by 35%.

PROJECTS

TaskPilot – Intelligent Task Planning Agent

- Engineered a multi-agent AI system with LangGraph and GPT-4, automating scheduling, prioritization, and time-blocking, boosting productivity by 50% for test users.
- Built retrieval-augmented memory with ChromaDB, preserving task context across sessions and enabling dynamic rescheduling with adaptive recommendations.
- Developed Streamlit dashboards and FastAPI modular endpoints, providing real-time visualization of tasks, reminders, and progress while ensuring scalable deployment and seamless API integration.

GenAI Customer Support Assistant

- Engineered a multi-modal conversational AI with GPT-4, LangChain, and Whisper, integrating text and speech to automate responses, reducing ticket resolution time by 40%.
- Implemented RAG pipelines with ChromaDB and FAISS embeddings to retrieve relevant knowledge from manuals and historical tickets, improving response accuracy by 35%.
- Developed a Streamlit dashboard for agents to track AI recommendations, monitor feedback, and fine-tune interactions, increasing adoption and customer satisfaction.

Personalized Content Generation Engine

- Designed an AI-driven content recommendation system using GPT-4, LangGraph, and Hugging Face Transformers, producing personalized marketing outputs, boosting engagement rates by 28%.
- Applied vector retrieval with Milvus embeddings to align content with user preferences, improving relevance and click-through rates by 25%.
- Integrated automated generation APIs via FastAPI, streamlining content workflows and reducing manual creation effort by 60% for enterprise-scale deployment.

EDUCATION

Master of Science in Information Systems **Aug 2023 - May 2025**
Central Michigan University | USA

Bachelor of Technology in Electronics & Communication Engineering **Aug 2018 - Jun 2022**
Annamacharya University | India

CERTIFICATIONS

- BM Generative AI Engineering Professional Certificate
- Generative AI Engineering with LLMs Specialization - **Coursera**
- AWS Generative AI Applications Professional Certificate
- AI Engineer for Developers Associate Certification - **LinkedIn Learning**
- Generative AI Leader Professional Certificate - **Coursera**
- Generative AI Concepts - **LinkedIn Learning**