

Muhammed Aflah

Madison, WI | (608) 440-4645 | aflah@wisc.edu | linkedin.com/in/muaf

Professional Summary

AI Engineer and Data Scientist specializing in LLM-powered systems, Retrieval-Augmented Generation (RAG), and scalable AI infrastructure. Experienced in building production-grade conversational AI, vector search pipelines, and intelligent data-driven applications using LangChain, OpenAI APIs, and Pinecone. Strong background in machine learning, distributed data systems, and end-to-end AI product development.

Education

University of Wisconsin–Madison

Master of Science in Information Science (Data Science)

Sep 2024 – May 2026

GPA: 3.6/4.0

Technical Skills

AI/LLM Engineering: Large Language Models (LLMs), RAG, Prompt Engineering, Embeddings, Semantic Search, Conversational AI

Frameworks & Tools: LangChain, OpenAI API, n8n, Hugging Face, Streamlit, Scikit-learn

Vector Databases: Pinecone, Embedding Pipelines, Similarity Search

Languages: Python, SQL, R

Cloud & Big Data: AWS, Azure, Google BigQuery, Hadoop, Hive, PySpark

MLOps & Systems: Model Deployment, API Integration, Scalable AI Systems, ETL Pipelines

Analytics & Visualization: A/B Testing, Statistical Analysis, Tableau, Power BI, Advanced Excel

AI Projects

LLM-Powered AI Assistant | RAG, LangChain, OpenAI, Pinecone

Jan 2026 – Present

- Architected a production-grade AI assistant leveraging LLMs with Retrieval-Augmented Generation (RAG) for context-aware responses.
- Built an end-to-end semantic retrieval pipeline including document chunking, embedding generation, and vector indexing.
- Integrated Pinecone vector database enabling low-latency similarity search over custom knowledge bases.
- Engineered advanced system prompts and prompt orchestration to control tone, persona, and recruiter-focused AI interactions.
- Designed modular Python backend for scalable API integration, knowledge base ingestion, and conversational inference.
- Reduced hallucinations and improved factual accuracy through grounding, temperature tuning, and response optimization.

MCP-Powered Voice Agent (Local LLM + Tool-Oriented Architecture)

Jan 2026 - Present

- Engineered a fully local, end-to-end voice assistant using Model Context Protocol (MCP) with Whisper-based speech transcription, LangChain orchestration, and Ollama-hosted Llama 3.2 for offline inference.
- Designed modular MCP-compatible tools including a SQLite database query engine and a DuckDuckGo web search fallback, enabling intelligent tool selection based on query context.
- Built an audio pipeline using PyAudio and local Whisper models to capture and transcribe real-time microphone input with zero cloud dependency and enhanced privacy.
- Developed a responsive desktop GUI using PyQt5 with QThread-based background workers to handle transcription and agent reasoning without UI freezing.
- Implemented structured tool-calling workflows and prompt routing to prioritize SQL database responses before dynamically triggering web search synthesis.
- Resolved critical environment and dependency issues including Python 3.14 incompatibility with pydantic-core, and configured system-level libraries (ffmpeg, portaudio) for stable audio processing on macOS.

Interactive ML Knowledge Bot | NLP, LangChain, Semantic Search

Sep 2023 – Oct 2023

- Developed an intelligent Q&A chatbot using LangChain, OpenAI API, and web-scraped knowledge sources.
- Implemented semantic search using Pinecone embeddings, improving retrieval speed and answer relevance by 50%.
- Deployed the application using Streamlit for real-time interactive querying.

Amazon Reviews NLP Analysis | NLP, Text Mining, ML

Sep 2024 – Dec 2024

- Performed large-scale sentiment analysis on 100,000+ Amazon reviews using NLP and machine learning models.
- Utilized NLTK and spaCy for text preprocessing, feature engineering, and classification achieving 85% accuracy.

Experience

Wayfair

Remote

AI Extern

Oct 27, 2025 – Dec 22, 2025

- Built multiple AI agents using n8n to automate workflows for trend detection, competitor tracking, and marketing content generation across the home goods domain.
- Conducted in-depth trend analysis on consumer demand, style preferences, and product narratives by analyzing market data, competitor launches, pricing strategies, and campaign signals.
- Designed automated data pipelines to aggregate agent outputs and transform unstructured signals into structured insights for business analysis.
- Consolidated AI-generated outputs into a live-updating Google Sheets dashboard, integrating trend signals, competitive benchmarks, and analytical insights to support category-level decision-making.

SuccessWorks, College of Letters and Science

Aug 2024 – Jun 2025

Data Assessment and Reporting Intern

Madison, WI

- Managed and secured sensitive datasets for 40,000+ students using structured data governance and cleaning pipelines.
- Developed automated dashboards in Tableau and Power BI, reducing manual reporting time by 40%.
- Performed advanced data analysis and post-event analytics to improve program decision-making by 25%.

iQuanti

Jun 2022 – Jul 2024

Senior Analyst

Bangalore, India

- Designed machine learning models for campaign performance and customer segmentation, improving targeting accuracy by 35%.
- Engineered scalable data lakehouse architecture using Python, PySpark, SQL, Hadoop, and Azure HDInsight.
- Implemented Bayesian Media Mix Models and forecasting systems, increasing ROI by 20% and reducing costs by 30%.
- Executed large-scale A/B testing frameworks that improved conversion rates by 15% and increased annual revenue by \$0.5M.