# Mahadev Mahesh Maitri

+1 (302) 685-6915 | mahadev@maitri.pro | linkedin.com/in/mahadev-maitri | github.com/mahadev9 | maitri.pro

## Technical Skills

**Programming Languages**: Python, JavaScript, TypeScript, C/C++, Kotlin, HTML, CSS, SQL
**Frameworks**: FastAPI, Django, Node.js, React, Bootstrap, Redux, Next.js, TailwindCSS, Flutter
**Cloud & Databases**: Amazon Web Services (EC2, ECR, EKS, Bedrock, MemoryDB, EBS, EFS, S3, SQS, Lambda, DynamoDB, RDS ), Google Cloud, MongoDB, PostgreSQL, Google Firebase, Docker, Kubernetes, Redis, Kafka
**ML Libraries**: PyTorch, TensorFlow, HuggingFace (Transformers, peft), LangChain, LangGraph, MCP (Model Context Protocol), Keras, Scikit-Learn, Gymnasium

## Professional Experience

**Fulcrum Digital** — New York City, NY
*AI Engineer* — October 2024 – Present

- **Architected** an enterprise document intelligence solution achieving **92% average extraction confidence** and **40% reduction in manual data entry**, by designing Doxtract-IDP using AWS Bedrock multi-modal models (Claude Sonnet 4) with a serverless Lambda-SQS-DynamoDB event-driven architecture.
- **Implemented** production-grade document processing with **90%+ field-level extraction accuracy** via human-in-the-loop validation and post-processing rules, deployed on serverless AWS infrastructure handling **60–600 second** processing windows per document.
- **Designed and deployed** high-throughput microservices on **AWS EKS** with horizontal pod autoscaling, load balancing, and CI/CD pipelines, delivering **99.8% system availability** while handling **300+ concurrent requests** during peak usage.
- **Architected** an event-driven GenAI pipeline using **LangGraph agent systems** with persistent memory and chain-of-thought reasoning, improving response relevance scores by **20%** and reducing manual policy review time by **15%**.
- **Engineered** a custom vector database retriever with **LLM-based reranking**, increasing search relevance scores by **25%** and achieving **87% precision** across **2,000+ synthetic insurance risk scenarios**.
- **Optimized** GPU infrastructure costs by migrating from local Ollama deployments to **AWS Bedrock API integration**, eliminating on-premise GPU dependencies and reducing monthly EC2 spend by **60%**.
- **Deployed** a self-hosted OCR service using **Vision Transformer (Phi3.5 Vision)** on EKS with **Ray clusters** for distributed computing, reducing monthly infrastructure costs by **20%**.
- **Built** a fault-tolerant RAG pipeline with **Kafka-based stream ingestion** and **Redis caching**, implementing custom retrieval agents and structured claim pattern analysis to achieve **87% precision** in insurance risk assessment across **2,000+ scenarios**.

**Department of Mechanical Engineering - University of Delaware** — Newark, DE
*Research Assistant* — January 2024 – February 2025

- Compiled a cross-platform application utilizing the Google Maps API to provide timely alerts about intersections based on live signal timings, enhancing navigation decision-making.
- Designed an algorithm to identify approaching intersections within a specified radius under 50ms, displaying signal data from an MQTT server corresponding to the upcoming signal phase using map data.
- Implemented an alert system using a dilemma zone algorithm to warn users 8 seconds before approaching intersections, utilizing real-time traffic data to avoid crossing red signals and enhance road safety.

**Optum - UnitedHealth Group** — Bengaluru, India
*Software Engineer* — July 2020 – July 2022

- **Designed** ETL pipelines using Python and AWS Glue to migrate **100,000+ records** with **98% accuracy** from legacy systems to AWS Redshift, ensuring seamless data transfer and centralized warehousing.
- **Engineered** NLP pipelines using SpaCy and NLTK to parse eligibility criteria and applicant details from **10,000+ forms** automatically, improving data extraction accuracy by **95%** with predictive models deployed on AWS SageMaker.
- **Integrated** OCR using Tesseract to automate data entry from **5,000+ scanned forms**, increasing document processing efficiency by **85%** and accelerating downstream workflows.
- **Built** low-latency REST APIs using FastAPI, integrating NLP and OCR models with PyTorch deep learning for real-time eligibility assessments, achieving **95% accuracy** in incoming application processing.
- **Executed** functional regression testing with Robot Framework and performance testing with JMeter, ensuring **97% SLA compliance** for RESTful APIs in the OMMS project through bi-weekly test suite execution and analysis.

## Projects

**Identifying Student Misconceptions in Math with Fine-Tuned LLMs**
*Python, Transformers, PEFT, BitsAndBytes* — July 2025 – October 2025

- **Employed** advanced parameter-efficient fine-tuning (PEFT) techniques like QLoRA with 4-bit quantization and Out-of-Fold (OOF) training to fine-tune LLMs including DeepSeek, Gemma-2, and Qwen3, maximizing model performance and robustness on imbalanced data.
- **Achieved** a final Mean Average Precision (MAP@3) score of **0.948** by engineering an ensembling pipeline that aggregated model outputs through weighted averaging and a custom disagreement-handling algorithm, leveraging the strengths of diverse architectures.

**Formula 1 Race Strategy Optimization with Deep Reinforcement Learning**
*Python, PyTorch, Gymnasium* — February 2024 – May 2024

- Designed and trained a Deep Q-Learning agent to optimize race strategies in Formula 1 by creating a data-driven environment using real-world racing data from FastF1.
- Developed a Markov Decision Process (MDP) with a reward system that considers pit stop penalties, tire wear, and lap times, enabling the agent to make strategic decisions about pit stops and tire selection to achieve optimal race performance.

## Education

**University of Delaware** — Newark, DE
*Master of Science in Computer Science - GPA: 4.0* — Aug 2022 – May 2024
*Relevant Courses: Algorithm Design, Advanced Deep Learning, Network Analysis, Compiler Construction*