[redacted]

[redacted]

AI/ML Engineer

☎ [redacted] @ [redacted]@gmail.com
🔗 linkedin.com/in/[redacted]/ 📍 [redacted]

SUMMARY

AI/ML Engineer with 3+ years of experience designing and deploying scalable, real-time AI systems across cloud-native environments. Specialized in large language models (LLMs), agentic AI workflows, and multimodal pipelines using LangChain, MCP and Hugging Face. Skilled in prompt engineering and vector databases for robust, low-latency deployments. Proven track record leading cross-functional teams, optimizing model throughput, and building autonomous, explainable systems for next-generation applications. Actively driving toward innovations in multi-agent AI, orchestration, and applied GenAI for intelligent automation, search, and decision systems.

SKILLS

AI Agents · Agent Orchestration · Autonomous Workflows · Multimodel Orchestration · Scalable AI Systems · Machine Learning (RNNs · CNNs · Transformers) · TensorFlow · PyTorch · Hugging Face · Prompt Engineering · Finetuning LLMs · Retrieval-Augmented Generation (RAG) · Ethical AI (GDPR · AI Act) · Python · NumPy · Pandas · Scikit-learn · AWS · Azure · GCP · Kubernetes · Docker · CI/CD · Statistics · RAGAS

CERTIFICATIONS

Azure AI Engineer Associate

Microsoft

Azure AI Fundamentals

Microsoft

EXPERIENCE

Technical Lead & ML Engineer

01/2025 - Present

[redacted] [redacted]

- Led a 7-member cross-functional team in designing a mobile GenAI app with real-time, personalized outputs, enhancing patient interaction through MCP workflows and LangChain orchestration.
- Developed an inference pipeline with ElevenLabs and Sync Labs for patient audiovisual generation, delivering high-quality, real-time outputs.
- Optimized model latency by 40%, reducing processing time from 91s to 55s, enabling near real-time edge device performance.
- Implemented API-based modular deployments using Docker and GCP Firebase, ensuring scalability and robust real-time model deployment on edge devices.
- Led design, deployed microservices, coordinated stakeholders; ensured scalable, robust execution.
- Conducted comprehensive evaluations on time-to-generation and response accuracy, maintaining consistent high-quality user experience.
- Directed agile development sprints, aligned with patient onboarding milestones, prioritizing audio-first development for trial phases.

Full-Stack ML Engineer

08/2024 - 01/2025

Toxicity AI [redacted]

- Designed and implemented an end-to-end GenAI pipeline using Mistral-4B, achieving 70%+ accuracy in identifying drug toxicity and overdose risks from synthetic patient notes.
- Engineered an advanced prompt strategy using few-shot learning and diagnostic prioritization to improve model precision in detecting adverse events.
- Created and transformed a synthetic dataset into realistic patient presentations with GPT-based prompts, enhancing data realism for training.
- Integrated FAISS vector search with LLM to enable real-time evidence retrieval, contributing to a RAG-based system.
- Utilized Hugging Face and TensorFlow for LLM deployment, integrating quantization to shorten training cycles and improve model performance metrics.
- Significantly reduced training time per epoch by migrating to A100 GPUs, accelerating hyperparameter tuning and boosting accuracy by over 10%.
- Led full system development from inception to deployment, including data design, model selection, and infrastructure setup for an LLM-driven toxicity detection proof-of-concept.

ML Engineer

06/2022 - 08/2024

[redacted] [redacted]

- Developed a dual-branch AI system integrating Vision Transformer (ViT) and TabTransformer, achieving 81% recall and 85% specificity on new patient data, comparable to finance performance monitoring.
- Engineered a custom sampling pipeline processing multi-million point EEG time-series into compressed formats, supporting advanced image-based modeling.
- Preprocessed and modeled multimodal data for a risk detection framework, demonstrating transferable skills in finance structured and temporal data domains.
- Built a high-dimensional transformer pipeline with 16×16 image patching and 7096-dim token embeddings for sequence modeling and representation learning.
- Solely architected, trained, tuned, and evaluated models using PyTorch, Hugging Face Transformers, and Weights & Biases, enabling robust experiment tracking.
- Produced a web-integrated, production-ready tool, illustrating infrastructure development and deployment expertise.