

SUMMARY

A recent Computer Science graduate with 3+ years of experience designing and deploying scalable, cloud-native infrastructure and applications. Proven expertise in full-stack engineering, modern DevOps methodologies, and architecting production-grade generative AI solutions. Skilled in building and orchestrating AI systems and agentic workflows that leverage LLMs. Proficient in using Infrastructure as Code (Terraform) and container orchestration (Kubernetes) to manage the complete lifecycle of these complex applications on AWS.

SKILLS SUMMARY

- **Cloud & DevOps:** AWS (EKS, EC2, S3, RDS, Lambda, VPC, IAM), Terraform, Kubernetes, Docker, Jenkins, CI/CD, Git
- **LLMs & AI Infrastructure:** RAG, LangChain, Docling, Pinecone, Supabase, Vector Databases, MCPs, Agentic Infra
- **Machine Learning:** PyTorch, TensorFlow, Scikit-learn, HuggingFace, Reinforcement Learning (DQN), Random Forests
- **Backend:** Node.js, FastAPI, Python, Java, PostgreSQL, MySQL, RESTful APIs, Bash Scripting
- **Product:** React, Next.js, TypeScript, HTML/CSS/JS, Agile Dev, Jira, Figma, Client Communication, UI/UX Design
- **Certifications:** Hugging Face (Fundamentals of Agents, 2025)

EXPERIENCE

- Freelance Web Developer (4 Months)**  
*UI/UX Designer & Developer*

Remote  
May 2025 – Present

  - Delivered 4 custom websites for clients in coaching, real estate, and education using modern stacks; used AI tools for content generation, layout prototyping, and SEO optimization, helping clients generate 100+ leads.
  - Managed the end-to-end product lifecycle, including client communication, timeline execution, and deployment.
- Mindgrasp AI (7 Months)**  
*R&D AI Engineer*

Remote  
December 2023 – July 2024

  - Built and shipped 6 Langchain-powered domain-specific agents with GPT-3.5, Med-PaLM 2, and Wolfram Alpha; integrated voice-based note-taking using AWS Transcribe and deployed product for customer use.
  - Developed a RAG-based internal Q&A assistant using Pinecone (2M+ vectors) to support operational teams; handled data ingestion, vector schema design, and prompt engineering (few-shot prompts with variables).
- Fidelity Investments (2 Summers)**  
*Software Engineer Intern*

Durham, NC — Jersey City, NJ  
June 2022 – August 2023

  - Cut EC2 and job runtime costs by 60% via restructuring of AWS resources (EC2, S3, Lambda), optimized cron jobs, and improved Jenkins pipelines; built Power BI visualizations with data parsed from JSON APIs.
  - Trained a Random Forest model on internal mutual fund data to analyze performance patterns, with data preprocessing in Python (Pandas) and visualizations of benchmarks to assess accuracy and feature importance.

PROJECTS

- Enterprise-Grade EKS Deployment for Node.js Application**

  - Automated AWS infrastructure provisioning using Terraform, deploying VPC, EKS cluster, EC2, RDS MySQL, S3, and IAM roles with cost optimization via NAT gateway elimination
  - Containerized Node.js application with Docker and deployed via Kubernetes manifests, implementing zero-downtime rolling updates and automated CI/CD pipeline using GitHub Actions.
- Clinical Document Fact-Checking System (Multimodal RAG Pipeline)**

  - Built multimodal RAG system processing 8 clinical research papers with images/tables into Milvus vector database using Docling, gpt-4o, and embeddings, achieving 75%+ accuracy in pharmaceutical claim validation
  - Implemented LLM-powered document processing pipeline with table repair, multi-page merging, and image captioning, plus semantic chunking and evidence filtering for automated claim-to-evidence matching
- DeepSeek Agentic RAG Crawler**

  - Engineered a conversational Agentic RAG system using GPT-4o-mini for natural language interactions with DeepSeek’s API Docs, integrated with Supabase for scalable vector storage of crawled web data.
  - Implemented parallel website crawling via crawl4ai, optimizing resource allocation to reduce compute costs by 80% while maintaining high-throughput data ingestion.
- Multilingual Voice Translation Application**

  - Engineered an AI-driven app using GPT-4o-mini and voice synthesis tools to deliver context-aware translations across 7 languages, deployed on Azure Web Services.
- Deep Q-Learning Tetris Agent**

  - Trained a PyTorch-based reinforcement learning agent using Deep Q-Networks (DQN), achieving autonomous gameplay capable of clearing over 1,000+ lines.

EDUCATION

- University of Maryland**  
*B.S. Computer Science*

College Park, MD  
2021 - 2025

  - **Relevant Courses:** Artificial Intelligence, Machine Learning, Computer Vision, Computer Systems, Compilers, Cryptography, Discrete Structures, Software Engineering, OOP, Data Structures & Algorithms
  - **Programs:** QUEST Honors, First Year Innovation and Research Experience (FIRE), Cybersecurity Club