Laurel, MD
+1(860)280-6380
edsonaguiar17@gmail.com

Edson Zandamela

Senior GenAI/DevOps Engineer with expertise in Python, AWS, GenAI, Life Sciences, and Infrastructure Automation.

Personal Website
LinkedIn
GitHub

PROFESSIONAL SUMMARY

Senior AI Infrastructure Engineer with 5+ years scaling **GPU** and **ML/GenAI platforms (AWS, Kubernetes, Terraform, EKS)** for **LLM** workloads across **biotech** and **big tech**. Architected **RAG systems** improving **R&D efficiency by 30%**, **optimized GPU infrastructure saving $1.2M+ annually**, and automated ML pipelines processing large datasets. Combines deep **cloud infrastructure expertise** with **hands-on AI development (LLMs, fine-tuning, multi-agent systems)**. Bilingual **(English/Portuguese)** with proven track record **building developer-friendly automation** and driving measurable business impact.

EDUCATION

University of The Cumberland's – MSc in Information Technology *August 2024*
Georgia Institute of Technology – MSc in Computer Science (ML Specialization) *May 2023*
Trinity College – BSc in Computer Science & Psychology (double major) *May 2020*

SELECTED AI CERTIFICATIONS

Google Cloud – AI Infrastructure: Introduction to AI Hypercomputer! *October 2025*
DeepLearning.AI – Claude Code: A Highly Agentic Coding Assistant! *December 2024*
LangChain – Project: Deep Research with LangGraph *July 2025*
Udacity – Foundations of Generative AI *Feb 2025*
LangChain – Project: Deep Research with LangGraph *July 2025*
NVIDIA – Augment your LLM Using Retrieval Augmented Generation *January 2025*
DeepLearning.AI – Finetuning Large Language Models *December 2024*
DeepLearning.AI – LangChain Chat with Your Data *November 2024*
DeepLearning.AI – Building Multimodal Search and RAG *September 2024*
DeepLearning.AI – ChatGPT Prompt Engineering for Developers *December 2024*
IBM Cloud For The Enterprise *July 2024*
DeepLearning.AI – Practical Multi AI Agents and Advanced Use Cases with crewAI, *October 2024*
DeepLearning.AI – Red Teaming LLM Applications *June 2024*
DeepLearning.AI – Building Systems with the ChatGPT API, *June 2023*
DeepLearning.AI – Introducing Multimodal Llama 3.2, *June 2023*
DeepLearning.AI – Pair Programming with a Large Language Model *May 2024*
DeepLearning.AI – Building Generative AI Applications with Gradio *July 2023*
DeepLearning.AI – Serverless LLM Apps Amazon Bedrock *July 2023*
LinkedIn Learning – DevOps with AWS *June 2023*
LinkedIn Learning – Learning Terraform *September 2022*
LinkedIn Learning – DevOps Foundations: Site Reliability Engineering *September 2022*
Udacity – AWS Machine Learning Foundations *May 2020*
Udacity – Google IT Support Professional Certification *May 2020*

TECHNICAL SKILLS

Generative AI: LLMs (OpenAI, Anthropic Claude), Retrieval-Augmented Generation (RAG), LangChain, Vector Databases (Weaviate, Chroma), Multi-Agent Systems, Transformers, Fine-Tuning.

AI/NLP/ML: Large Language Models (OpenAI, Anthropic Claude, **Apple Foundation Models**), **Encoder/Decoder Architectures**, **Fine-Tuning**, Retrieval-Augmented Generation (RAG), **Knowledge Graphs**, **Natural Language Processing**, **Bias Detection & Mitigation**

Cloud & AI Infrastructure: AWS (EC2, S3, Lambda, EKS), Azure, Kubernetes, Terraform, Docker, Prometheus/Grafana GPU Infrastructure Optimization, Production ML Security, Cost Analytics & Monitoring.

Software Development: Python, JavaScript, REST APIs, FastAPI, Microservices, CI/CD (GitHub Actions, GitLab)

Tools/Frameworks: Hugging Face, PyTorch, Streamlit, Gradio, Helm, GitOps, MLflow

Web Development: Python, Gradio, Streamlit, HTML, CSS, JavaScript, HelmCharts, REST APIs, CI/CD pipelines

WORK EXPERIENCE

Apple Inc. – *Platform Infrastructure Engineer - (Contractor)* Cupe, CA, *April 2025 – Present*

- **GPU Infrastructure Cost Optimization:** Built a production AWS Lambda system processing 3**50K+ CloudWatch/CloudTrail events** daily to monitor GPU utilization trends **across 50+ AWS accounts.** Identified **$2M+ in annualized GPU savings'** costs and reduced manual reporting from **1 week - to just under 15 minutes** (97% improvement).
- **AI Platform Evaluation & Security:** Tested Apple's internal GenAI/Models; implemented IAM best-practices for **HyperPod clusters** with zero downtime.
- **Observability & Dashboards:** Created Datadog dashboards tracking GPU capacity, temperature mismatches, and power draw per host; automated alerting for anomalies in GPU health and utilization.
- **Kubernetes & HyperPod Engineering:** Built and upgraded HyperPod + EKS clusters, validated Cilium + CNI integrations, and deployed Ray job workloads for distributed ML training.
- **GPU Metrics Pipeline:** Enhanced the internal GPU metrics pipeline to use a time-series database backend for long-term **NVIDIA SMI metric** retention and cluster and account-level analytics.
- **Cross-Vendor Collaboration:** Partnered with AWS support to resolve GPU metric discrepancies and evaluated Run:AI for partial GPU allocation and job-level scheduling.

Arcaea – *DevOps Engineer* Boston, MA *August 2023 – March 2025*

- Architected an enterprise AI platform using **LLMs and RAG**, improving content production efficiency by **20%** through automated report generation and molecular data analysis.
- Deployed scalable **vector database solutions** (Weaviate/Chroma) integrated with LangChain, enhancing search accuracy for biotech R&D teams.
- Led cross-functional collaboration to design **multi-agent LLM systems**, reducing experiment turnaround time by **30%** via automated workflows.
- Built **cloud-native AI applications** on AWS EKS using Kubernetes and Terraform, ensuring **99.9% uptime** and reducing infrastructure costs by **90%**.
- Implemented real-time monitoring for AI systems using **Prometheus/Grafana**, enabling proactive incident resolution.

Anagenex – *DevOps Engineer* Lexington, MA *August 2021 – August 2023*

- Designed Flask-based internal tools to automate data-sharing workflows, reducing manual task time by 50% for cross-functional teams.
- Developed **ML pipelines** for drug discovery, processing **large biochemical data** on AWS, reducing pipeline failures by **25%** via robust error handling.
- Automated **CI/CD workflows** for ML model deployment using GitLab and Docker, accelerating model evaluation cycles by **40%**.
- Collaborated with data scientists to translate research requirements into production-grade AI solutions, improving experiment throughput by **35%**.

ZebiAI Therapeutics – *DevOps Associate* Waltham, MA *July 2020 – July 2021*

- Built Python/Flask microservices to streamline data ingestion for chem informaticians, significantly improving pipeline efficiency.

- Migrated **10TB+ datasets** post-acquisition with **100% integrity**, ensuring seamless integration with Relay Therapeutics.
- Secured Azure environments with IAM policies and Intune, minimizing compliance risks.

Trinity College – *IT Consultant* Hartford, CT *September 2016 – May 2020*

- Provided campus-wide technical support for hardware, software, and network infrastructure
- Configured and maintained computer labs and classroom technology
- Led network architecture training sessions for new employees
- Created and maintained technical documentation
- Managed video-conferencing equipment setup and troubleshooting
- Mentored trainees on IT systems debugging and problem-solving

PROJECT EXPERIENCE

AI Learning Hub LLC – Founder & Lead Instructor (2024 – Present) *(2024 – Present)*

- Founded bilingual (English/Portuguese) AI education company creating practical GenAI courses for developers across 4 continents (US, Europe, South America, Africa)
- Published "AI-Powered Content Creation & Automation" on Udemy (5★ rating); developing second course on building production-ready LLM tools with LLaMA 3.2
- Curate weekly AI newsletter analyzing emerging trends, tools, and best practices for subscribers
- Produce daily Portuguese-language content on LinkedIn reaching developers in Portuguese-speaking markets (Brazil, Portugal, Mozambique, Angola)

Girls Can Code Club – Founder, Maputo, Mozambique *2020 - 2023*

- Founded remote coding education initiative teaching web development and programming fundamentals to 50+ young women during COVID-19 pandemic
- Partnered with US Embassy Maputo to provide free STEM education access in underserved Mozambican communities
- Designed curriculum covering HTML, CSS, JavaScript, and Python basics; students built portfolio projects demonstrating learned skills

Health Interoperability – https://github.com/edsna/Health_Interoperability *Spring 2021*

- Developed a web app that allows patients to choose to share their medical records with any clinician.
- Used Smart on FHIR API to perform calls and store data into a Firebase database.