Alex Carter

📍USA | 📧 alex.carter@abcd.com | 🔗 linkedin.com/in/alexcarter-ai

SUMMARY

AI Engineer with 10+ years of technical experience spanning **data analytics, data science, and applied AI engineering**, including 3 years specializing in **deep learning model development, deployment, and optimization** across **NVIDIA DGX**, **Azure AI Foundry**, and hybrid **on-prem/cloud** environments. Passionate about building scalable AI systems, accelerating model performance with GPU-optimized workflows, and operationalizing LLMs, vision, and generative AI solutions in enterprise settings.

CORE SKILLS

- **AI Platforms:** NVIDIA DGX Spark / DGX Station, Azure AI Foundry, Azure ML, OpenAI Service, Hugging Face, Ollama, ONNX Runtime
- **Languages:** Python, SQL, Bash, PowerShell, YAML, TypeScript (basic)
- **Frameworks & Tools:** PyTorch, TensorFlow, Scikit-learn, LangChain, FastAPI, Docker, Kubernetes, GitHub Actions, MLflow
- **Cloud & Infrastructure:** Azure (VMs, Key Vault, AKS, Storage, ACR), Linux, NVIDIA NGC, REST APIs, Entra ID Authentication
- **AI Specialties:** LLM fine-tuning, computer vision, retrieval-augmented generation (RAG), model optimization, GPU profiling
- **Data Stack:** Azure Synapse, Databricks, Snowflake, Power BI, Pandas, NumPy, Spark

---

PROFESSIONAL EXPERIENCE

AI Engineer | ACME Analytics Group

Austin, TX | Mar 2022 – Present

- Designed and deployed **GPU-accelerated AI workloads on NVIDIA DGX Spark and Azure AI Foundry**, enabling a 40% reduction in model training time for computer vision and NLP projects.
- Built **LLM-based workflow agents** using Azure AI Foundry's orchestration tools for document intelligence and autonomous data summarization.
- Implemented **RAG pipelines** integrating Azure AI Search, Cognitive Services, and custom embeddings for enterprise knowledge retrieval.

- Containerized Python-based inference services using **Docker and ACR**, deploying across **on-prem DGX clusters and Azure Kubernetes Service (AKS)**.
- Developed internal monitoring tools leveraging **Prometheus, Grafana, and NVIDIA DCGM** to track GPU utilization and model performance.
- Collaborated with data scientists to refactor legacy TensorFlow models into **PyTorch/ONNX** for better inference throughput on A100 GPUs.

Key Technologies: Python, PyTorch, Azure AI Foundry, FastAPI, DGX Spark, Azure ML, LangChain, MLflow, Docker

---

Senior Data Scientist | ACME Insights

Dallas, TX | Jun 2017 – Feb 2022

- Delivered end-to-end **predictive analytics and machine learning solutions** for Fortune 500 clients using Azure ML and Databricks.
- Developed production-grade data pipelines (ETL/ELT) and implemented **model versioning and CI/CD** for analytics models.
- Partnered with engineers to integrate AI model outputs into client-facing dashboards and APIs.
- Mentored junior analysts and contributed to the team's transition from analytics-focused workflows to **AI-first development**.

Key Projects:

- Customer churn prediction (XGBoost + Azure Data Factory)
- Product demand forecasting (Prophet + Databricks Delta Lake)
- Sentiment classification (Azure Text Analytics + Scikit-learn)

Data Analyst | ACME Solutions

Houston, TX | Apr 2014 – May 2017

- Built data models, reports, and dashboards using SQL, Power BI, and Python automation scripts.
- Led data migration from on-prem SQL Server to Azure Synapse, introducing pipeline automation and governance processes.
- Recognized for developing KPI dashboards used by executive teams to drive business insights.

EDUCATION

M.S. in Data Science — University of Texas at Austin

B.S. in Computer Information Systems — Texas State University

---

CERTIFICATIONS

- NVIDIA Certified AI Specialist
- Microsoft Certified: Azure AI Engineer Associate
- Microsoft Certified: Azure Solutions Architect Expert
- TensorFlow Developer Certificate

SELECT PROJECT HIGHLIGHTS

- **Hybrid AI Deployment Platform:** Built a unified deployment framework for vision and LLM models across DGX (on-prem) and Azure (cloud).
- **LLM Fine-tuning for Legal Document Summarization:** Used NVIDIA NeMo and Azure AI Foundry to fine-tune Llama-3 models, improving summarization accuracy by 22%.
- **AI Model Observatory:** Developed GPU utilization dashboards and inference performance benchmarks for multi-model serving pipelines.