Adnan Ashraf

(419) 320-1540|aaadnan259@gmail.com|https://github.com/aaadnan259|adnanashraf.dev|Perrysburg, Ohio

SUMMARY

Early-career AI / Machine Learning Engineer with hands-on experience building production-ready Generative AI systems, AI services, and ML-powered applications. Proven ability to design Python-based microservices, RAG-style AI pipelines, and internal APIs that translate machine learning outputs into measurable business value. Seeking an AI/ML Engineer role contributing to firmwide AI platforms, scalable AI services, and production GenAI integrations.

EDUCATION

University of Toledo **Graduated: December 2025**

Bachelor of Science in Computer Science and Engineering Technology *Toledo, OH*

WORK EXPERIENCE

Code Echo **July 2024 – Present**

Software Engineering Intern *Toledo, OH*

- Built and maintained production Python-based backend services and REST APIs supporting client applications for retail and real estate businesses (including Wild Wings and Things and Rocky Ridge Development), enabling reliable workflows for 50+ daily active users per system.
- Delivered 2+ production web applications, modernizing digital presence and internal workflows, directly supporting customer engagement and revenue operations for regional SMB clients.
- Engineered and debugged Python, Java, and C modules across active client projects, reducing post-deployment defects by ~15%, lowering support overhead, and improving system stability for live customer-facing platforms.
- Implemented AI-augmented automation and NLP-driven tools within internal and client-facing systems, reducing repetitive operational tasks by ~30%, strengthening Code Echo's positioning as an AI-capable software consulting partner.
- Collaborated within a distributed Agile team of 3–5 engineers to deliver concurrent client projects on time, maintaining 100% on-time sprint completion, ensuring predictable delivery timelines and consistent client satisfaction.
- Planned and executed a zero-downtime IT and network migration for Toledo Recycle Services (supporting 85–90 employees), completing the transition with 0 minutes of downtime and preserving uninterrupted business operations.

PROJECTS

EchoBot - Generative AI & NLP Virtual Assistant|echobot-sics.onrender.com

Tech Stack: Python, Generative AI, NLP, LangChain, Vector Search, FastAPI, Render

- Architected a real-time AI microservice exposing 5 REST & WebSocket endpoints, implementing 9+ async functions to handle 100+ concurrent user sessions without blocking the main event loop.
- Built a Top-K (k=3) RAG pipeline using ChromaDB cosine similarity, dynamically injecting relevant memory chunks into Google Gemini's 1M+ token context window, achieving <200ms average response latency.
- Engineered graceful degradation patterns for external integrations (VoiceEngine, VectorDB), ensuring continuous system availability and preventing crashes during downstream service failures.

IntelliStock - AI Powered Inventory Management System

Tech Stack: Python, Machine Learning, Django, PostgreSQL, scikit-learn

- Developed a Linear Regression forecasting engine using scikit-learn that analyzed historical sales trends to predict demand, improving forecast accuracy from 65% → 85% (R^2 score > 0.8).
- Architected a scalable data pipeline capable of handling millions of sales records by offloading temporal aggregation to PostgreSQL (via TruncDay), preventing application-level memory bottlenecks.
- Optimized database performance using Django ORM's select_related, reducing query overhead by ~98% (solving N+1 query issues), directly contributing to a 20% reduction in stockouts for simulated users.

Generative Metaball Engine - Senior Design Capstone (Award Winner)
Tech Stack: Python, NumPy, OpenCV, Systems Optimization

- Achieved ~100x performance speedup by replacing standard loops with NumPy vectorization and NumExpr, utilizing CPU SIMD instructions for real-time pixel processing.
- Architected a multi-threaded producer-consumer pipeline with queue-based backpressure, decoupling rendering from I/O to prevent memory leaks and ensure zero frame drops.
- Optimized memory management using pre-allocated contiguous arrays, effectively eliminating Garbage Collection (GC) pauses to maintain smooth 24/7 autonomous operation.
- Secured 3rd Place Award at the Senior Design Expo, recognized by judges for technical complexity in O(P×N) algorithm optimization and system stability.

TECHNICAL SKILLS

- **Programming Languages:** Python, Java, JavaScript, C, SQL, HTML/CSS
- **AI / Machine Learning:** Large Language Models (LLMs), Prompt Engineering, Generative AI, RAG, NLP, Predictive Modeling, Model Evaluation (RMSE, R^2)
- **Frameworks & Libraries:** PyTorch, scikit-learn, pandas, NumPy, spaCy, NLTK, LangChain, FastAPI
- **AI Systems & Infrastructure:** REST APIs, Microservices Architecture, Vector Search (ChromaDB/FAISS), AI Model Deployment, Cloud ML Services, API Orchestration, Vector Embeddings, Scalable AI Systems, Production GenAI Pipelines
- **Tools & Platforms:** Git, GitHub, AWS, Render, PostgreSQL, Django, React, Jira, Docker