.Jacob Furtaw

Jfurtaw97@gmail.com | 410-533-7663 | www.linkedin.com/in/jacob-furtaw/ | ***www.jfcoded.com/projects***
Baltimore, MD | Willing and Ready to Relocate Anywhere

Professional Summary

Machine Learning Engineer with four years of experience across academia and professional experience in AI-driven research and full-stack software development, specializing in AI Agent development, Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), and wrangling and pre-processing large datasets. Proficient in rapid prototyping of local, scalable, production-grade AI applications. Adept at collaborating with cross-functional teams to deliver data-driven solutions aligned with business goals, with a proven track record of optimizing workflows and enhancing system performance.

Skills

Programming Languages: Python, JavaScript; Familiar With: C++
Python Libraries/Frameworks: Llama-Index, Langchain, FastAPI, Django, Pandas, HuggingFace, **PyTorch**, Transformers, **Scikit-Learn**, Accelerate, RAPIDS
Tools & Platforms: Git, **Docker**, Gitlab, Jupyter Notebook, Ollama, Huggingface, NVIDIA AI Foundation Models(build.nvidia.com), Mistral Vibe, Claude Code
Cloud: Currently completing official AWS “Machine Learning Engineering for Production (MLOps)” specialization

Work Experience

Machine Learning Engineer | SurgePoint Software (Stealth Startup) | Remote August 2023 - April 2025

- Utilizing **data engineering** skills to reduce 200 million lines of unstructured data into a 13-million-line structured dataset, increasing semantic relevance scores by 50-75%, and reducing model hallucinations
- Developed custom data wrangling and cleaning techniques for large-scale datasets, ensuring data integrity and enabling exploratory analysis for actionable business insights
- Designed, rigorously tested, and implemented a complex **Retrieval-Augmented Generation** (RAG) pipeline that uses a vector database (Milvus, ChromaDB) to supply various LLMs with my custom dataset
- Collaborated with a 6-person cross-functional startup team in weekly standups and sprint reviews, delivering actionable insights and aligning technical efforts with business goals

Advanced Repair Agent | Geek Squad | On-Site | Seasonal March 2022 - Present

- Designing operational improvements alongside new management that increased the team's productivity by over 50% and earned me a letter of recommendation from upper management
- Consistently ranked the **top performing** Advanced Repair Agent across our marketplace, resolving hundreds of hardware and software repairs monthly across diverse devices and operating systems

Research Projects

Cloak AI | Closed Source, but happy to demo it!

- Using NVIDIA's Nemotron Nano 3 Hybrid Mamba 2/Transformer MOE model to build a general-purpose **AI Agent**, with **multi-tool access and a custom prompt** driving the model to provide the best answers to the user's query.
- Built and designed a custom UI using React that rivals modern web UIs from current Foundation model providers.
- Built and deployed custom tools for web search, stock tracking, and real-time access to news, sports scores, and weather, extending Nemotron's capabilities for dynamic user interactions.

Automatic Identification of Equivalent Mutants using an ASTNN(GNN) | Project Link

- Collaborated in a five-person Scrum team, participating in sprint planning, daily standups, and sprint reviews to deliver a transformer-based model (CodeBERT) for binary classification
- Optimized data preprocessing with custom Python parsers, tuned hyperparameters, and enhanced Jupyter notebook training scripts
- Improved F1 and accuracy scores from 79% to 92% from past researchers through dataset balancing techniques

Chat RAG | Project Link

- Created a RAG-powered chatbot with a Gradio user interface, supporting **local and API inference** from any of the hundreds of Ollama and HuggingFace models, as well as any models from OpenAI, Anthropic, and NVIDIA NIMS
- Engineered a modular Python architecture with 5+ features for model management, featuring dynamic model switching, custom prompt integration, model parameter tuning, quantization options, and many more

- Designed flexible data ingestion from three diverse sources (local files, GitHub repositories, and vector databases)

Education

Bachelor of Science in Computer Science, Software Engineering Concentration December 2023
Towson University, Towson, MD
Clubs: Machine Learning Research Group