Daniel Cho

AI Engineer

Seattle, WA · daniel.cho@example.com · (206) 555-0138 · LinkedIn · GitHub · Portfolio

SUMMARY

AI engineer with 5 years building and deploying machine learning and LLM-based systems in Python, PyTorch and AWS. Built a retrieval-augmented support assistant handling 60% of inbound tickets and a demand forecasting model that cut excess inventory by 18%. Comfortable owning a model from data pipeline through production monitoring.

EXPERIENCE

AI Engineer — Harborline Support Systems

Seattle, WA · Apr 2023 – Present

- Built a retrieval-augmented generation assistant over 40,000 help articles and past tickets, using OpenAI embeddings and a Pinecone index, now resolving 60% of inbound support tickets without a human agent.
- Fine-tuned a Llama 3 8B model on 25,000 labeled support replies to match company tone, cutting agent edit rate on drafted responses from 45% to 12%.
- Deployed the assistant behind a FastAPI service on AWS ECS with request logging and a fallback to a smaller model under load, holding p95 latency under 800ms at 3,000 requests a day.
- Added an automated eval suite of 300 scenario prompts that runs on every model or prompt change, catching 9 regressions before release over the past year.

Machine Learning Engineer — Cascadia Freight Analytics

Seattle, WA · Jul 2020 – Mar 2023

- Built a demand forecasting model in PyTorch on 3 years of shipment data across 200 routes, cutting excess warehouse inventory by 18%.
- Trained and served a route-delay classifier with scikit-learn and SageMaker, flagging 70% of late shipments a day ahead for dispatchers.
- Set up an MLflow pipeline tracking 400+ training runs, cutting the time to compare model versions from a day to under an hour.
- Migrated batch scoring jobs to a nightly Airflow DAG, reducing manual reruns from weekly to none over 6 months.

SKILLS

ML and LLMs: PyTorch, scikit-learn, Hugging Face Transformers, LangChain, prompt engineering, fine-tuning, RAG

Languages: Python, SQL, basic TypeScript

Infrastructure: AWS (SageMaker, ECS, Lambda), Docker, FastAPI, Airflow, MLflow

Data and vector stores: PostgreSQL, Pinecone, pandas, Spark (basic)

CERTIFICATIONS

AWS Certified Machine Learning – Specialty, 2022

EDUCATION

M.S. Computer Science, University of Washington, 2020

Fictional example by FindAJobAI: the person, employers and numbers are invented.