DANIEL LEON – AI/ML Infrastructure Engineer

🌎**Seattle, WA** 📞**253-592-8353** ✉️**dannylleon@gmail.com** 🌐**github.com/dleon86** 🔗**linkedin.com/in/daniel-leon-ai-swe**

PROFESSIONAL SUMMARY

AI/ML Engineer building production-scale **generative AI infrastructure** at Meta. Specializing in multimodal model evaluation systems, ML inference services, and large-scale data pipelines for video, audio, and image generation. Combines **full-stack engineering** (*React*, *FastAPI*, *Docker*) with deep **ML expertise** (*PyTorch*, computer vision, NLP) and a research background in applied mathematics and computational science. Track record of independently designing and shipping systems used by **thousands**.

TECHNICAL SKILLS

- **Languages:** Python, TypeScript, JavaScript, C++, SQL, MATLAB
- **ML / AI:** PyTorch, scikit-learn, XGBoost, OpenCV, LangChain, Pinecone, Hugging Face, RAG Pipelines, Fine-Tuning (BERT, CNNs), Embeddings, Agentic Workflows
- **Infrastructure:** Docker, FastAPI, React/Next.js, REST APIs, Gunicorn/Uvicorn, TLS/HTTPS, FFmpeg, yt-dlp, Git
- **Data:** Presto/Hive, PostgreSQL, PySpark, Pandas, NumPy, ETL Pipelines, Distributed Storage Systems, Data Validation
- **MLOps:** GPU Cluster Deployment (**H100**), Model Serving, Inference Pipeline Orchestration, CI/CD, Container Registries (ECR), Slurm
- **Visualization:** Plotly, Matplotlib, Dashboarding

PROFESSIONAL EXPERIENCE

Software Engineer 2 *(Scope: AI/ML Infrastructure Engineer)* | *July 2025 – Present*
Meta (via Tundra Technical Solutions) | *Seattle, WA*
Multimedia Vertical Generation Team

Evaluation Platform & Annotation Infrastructure

- Architected and co-developed a human evaluation platform used by **10,000+ annotators**, leading React/TypeScript frontend development and contributing to backend services, enabling scalable multimodal model assessment across video, image, and audio
- Improved annotation precision from *100ms* to **10ms** time resolution across task and audit UIs.
- Built a custom *React* annotation UI (**~1,400 lines**) for video character evaluation with features like reference frame capture and bounding-box crop tools.

ML Inference & Service Architecture

- Driving migration of ML inference services from legacy RPC (*Thrift*) to **Docker + FastAPI (HTTP)**, contributing to a standardized deployment pattern and executing multiple system transitions.
- Deployed a *ResNet18 2.5D* video classification model on **H100 GPU infrastructure** with sub-second cold start.
- Extended and applied configuration-driven inference pipelines processing **10M+ multimodal assets** for tasks including diarization and synthetic media generation.

Data Engineering & Pipeline Operations

- Built and operated an end-to-end media processing pipeline for dubbing/voice cloning evaluations, enabling scalable processing across **12+ languages** and supporting large-scale model validation workflows.
- Engineered ETL workflows ingesting **millions of records** into distributed databases,

DANIEL LEON – AI/ML Infrastructure Engineer

🌎**Seattle, WA** 📞**253-592-8353** ✉️**dannylleon@gmail.com** 🌐**github.com/dleon86** 🔗**linkedin.com/in/daniel-leon-ai-swe**

using complex *Presto/SQL* validation queries.
- Co-developed a data preservation pipeline for **114+ critical ML training tables**, handling schema validation and privacy compliance.
- Ranked in the **top tier of AI-assisted development adoption**, leveraging agent-based workflows and AI-generated diffs to accelerate engineering velocity and system iteration.

Independent AI/ML Engineer | *April 2024 – July 2025*
Seattle, WA

- Designed and built an AI-powered product ratings platform (Axiomatiq) that uses agentic LLM workflows (LangChain + OpenAI) to extract structured data from unstructured text, generate embeddings, and auto-populate relational databases. Deployed to two consumer verticals: artisanal foods (tastemongers.com) and chef's knives (edgemongers.com).
- Developed **RAG pipelines** with *LangChain* & *Pinecone* for domain-specific question answering; benchmarked retrieval accuracy and latency.
- Fine-tuned **BERT** for specialized NLP classification with limited labeled data, implementing rigorous cross-validation frameworks.

Data Scientist II | *August 2023 – April 2024*
Meta (via Vertisystem) | *Seattle, WA*

- Developed ML models to optimize wearable device ergonomics, achieving **98% population fit coverage**.
- Automated *Ansys* simulation workflows with *Python*, **reducing analyst iteration time by 60%**.
- Built real-time demographic monitoring dashboards (*Pandas, SQL, Plotly*).

Data Scientist / Research Engineer | *June 2018 – August 2023*
University of Washington, Lutz Research Group | *Seattle, WA*

- Led development of the **Data-Enhanced Diagnostics (DeDx)** model, improving molecular test sensitivity using ensemble ML methods.
- Trained **CNNs** to predict SARS-CoV-2 spike protein concentrations with **95% accuracy**.
- Designed experiments and statistical analyses for point-of-care diagnostics; developed *MATLAB* simulations.

EDUCATION

- **MS, Applied Mathematics** | *University of Washington (2023)*
- **BS, Chemical Engineering** | *University of Washington (2015)*

PROJECTS

Ratings Platform | github.com/dleon86/axiomatiq_ratings_db

Edgemongers | edgemongers.com/ratings

DANIEL LEON – AI/ML Infrastructure Engineer

🌎**Seattle, WA** 📞**253-592-8353** ✉️**dannylleon@gmail.com** 🌐**github.com/dleon86** 🔗**linkedin.com/in/daniel-leon-ai-swe**

Tastemongers | tastemongers.com/ratings

Built a generalized AI ratings platform that uses LLM agentic workflows (LangChain, OpenAI) to extract structured product data from unstructured sources, generate embeddings, and sync to consumer-facing sites. Deployed across two verticals: artisanal foods (cheeses) and chef's knives. Stack: Python, PostgreSQL/Neon, Next.js, Vercel, Amazon PA-API.

References available upon request