Aaron Leevord Fernandes

Data Scientist: Recommendation, Reinforcement Learning and AI Agents

+1 (930) 333-4174 | aaronferns38@outlook.com | aaronfernandes.dev
github.com/Aaronaferns | linkedin.com/in/aaron-leevord-fernandes

Summary

Builds recommendation and reinforcement learning models and AI agents, and the evaluation that shows whether they actually help. Works in Python, pandas and SQL on large data, including 2.8 million shopping events and a 2-million-image dataset. Spent about two years as a machine learning research engineer at Indiana University, after a year at TCS writing Java services and SQL for HDFC Bank. In recent projects, compares new models against strong baselines and reports confidence intervals.

Education

Indiana University Bloomington | Bloomington, IN | 2023 – May 2025
M.S. Intelligent Systems Engineering (Computer Engineering), GPA 3.9 of 4.0

Goa College of Engineering | Goa, India | 2017 – 2021
B.E. Electronics and Telecommunication Engineering, GPA 3.7 of 4.0

Projects

Next-Item Recommendation, personalized e-commerce recommender | *Python, PyTorch, XGBoost* 2026

- Compared 18 recommenders on 2.8 million RetailRocket shopping events, from popularity counts to transformers with reinforcement learning heads, including a new max-entropy variant. Each one ranked the full catalog of 42,000 items to predict the next item a shopper would interact with, scored by NDCG@10 on a held-out week. A new hybrid model, a transformer that corrects a simpler recommender, beat the strongest baseline by 14%, and a two-stage XGBoost reranker scored best overall.
- Broke the score down by user group. Shoppers going back to an item they had already viewed were 9% of users but supplied about 62% of the new model's score, and on items new to them it tied the other strong models. Swapping reinforcement learning objectives moved results by less than the bootstrap error bars, which puts the credit on the model design rather than on RL.

LedgerBench, evaluation benchmark for LLM agents | *Python, asyncio, Claude API* Oct 2026

- Wrote 40 ledger tasks for LLM agents, each with an answer computed twice by separate code, plus an LLM judge that grades one rubric item per call. Before any real model ran, tested the scoring with fake models: a perfect answer had to get full marks, and empty, hedged or prompt-injection answers had to get zero.
- Each agent is scored on pass@1 for one try, pass@k for any of k tries and pass^k for all k tries, along with cost per solved task and tool errors. Built Wilson, bootstrap and task-clustered confidence intervals, and ran paired permutation tests and power analysis on simulated models, which showed 40 tasks can only catch gaps of about 15 points. Claude Haiku and Sonnet both solved all 40 (true rate between 91% and 100%), Haiku at under half the cost per solved task, so the tasks need to be harder.

Computer-Use Automation System, agentic AI | *Python, Playwright, Claude API, SQLite* Sep – Oct 2026

- Taught an AI agent to do a task in an old web ERP once, in a real browser. The run is saved as a file a person reviews and approves, and from then on a plain interpreter replays it with no model calls.
- Because the task posts payments, the controls are strict. The model never sees real credentials, personal data stays out of the logs, and any action not on an approved list is blocked. A SQLite ledger keeps a payment from being posted twice, a lost reply goes to a person instead of being retried, and over 850 tests cover it.

NCAA Ticket-Sales Forecasting | *Python, scikit-learn, Tableau* 2024

- Forecast ticket sales for NCAA women's basketball with Random Forests on game, team and schedule data, where feature engineering did most of the work. Presented the results in Tableau to a non-technical judging panel and placed in the top 5 of a national analytics competition.

ScoutVideo, conversational AI voice agent | *Python, vLLM, LangGraph, FalkorDB* 2026

- A voice agent for asking questions about driving footage out loud, and it stops talking the moment someone cuts in. A Qwen3-8B router sends frame, video, comparison and search questions straight to a tool and everything else to the main model, and scene search uses RAG over a knowledge graph of driving scenes.

Re-Diffusion, data pipeline and profiling for a diffusion transformer | *PyTorch, DDP*

- Streamed about 2 million images through a data loader that ran far faster than the model could use them, then swapped in PyTorch's fused attention to more than double throughput and halve peak memory.

Experience

Indiana University Bloomington Apr 2024 – Jul 2026
Research Engineer *Bloomington, IN*

- Astemo Americas sponsored a SMART Lab benchmark of 8+ image-editing and video-generation models on nuScenes and Argoverse driving data. Wrote the PyTorch pipeline behind it, which prepares the data, runs every model across several GPUs and scores what comes out.
- Designed the geometry and realism metrics and tested them on real frames against LiDAR depth and 3D boxes before trusting them on generated ones. They became the benchmark's main comparison. Wrote up the metrics and failure cases for the sponsor and contributed experiments to a paper under review at WACV.
- Made custom PointMaze tasks in MuJoCo for a goal-conditioned reinforcement learning project. Adding hindsight experience replay, which lets a Soft Actor-Critic agent learn from goals it missed, raised task success from 15% to 55%, with experiments tracked in MLflow and Weights & Biases.
- At the Leung Research Group, trained a U-Net that stands in for a three-minute physics simulation and answers in under a second for one family of metasurface designs, and built a pipeline that turns time-series OCT videos of tissue into histology-equivalent images.

Heartland Community Network (CHOMP) Jan 2026 – Apr 2026
Software Engineer (contract) *Remote*

- Joined a small team before launch and turned product requirements into Node.js and Express services on AWS for recommendations, event-driven analytics and scheduled jobs. After go-live, added logging and monitoring and traced production issues to their root cause.

Tata Consultancy Services (client: HDFC Bank) Jan 2022 – Jan 2023
Software Engineer *India*

- Wrote the SQL, Java and Spring Boot microservices and REST APIs behind HDFC Bank's transaction processing, automated reconciliation in its batch jobs, and worked with QA and infrastructure teams on testing, releases and incidents.

Technical Skills

Data: Python, pandas, SQL, NumPy, scikit-learn, XGBoost, Tableau, feature engineering

Machine learning: recommender systems, two-stage ranking, SASRec, EASE, reinforcement learning, transformers, Random Forests, predictive modeling, PyTorch, DDP

Metrics and statistics: offline evaluation design, metric validation, NDCG, MRR, pass@k, bootstrap and Wilson confidence intervals, paired permutation tests, power analysis

GenAI: agentic AI, LLM agents, conversational AI, tool calling, RAG, LLM evaluation, LLM-as-judge, vLLM, LangGraph, Claude API

Engineering: Java, Spring Boot, AWS, Docker, Git, MLflow, Weights & Biases, SLURM, SQLite, Playwright

Certifications

AWS Certified Developer – Associate
AWS Certified Cloud Practitioner