Aryan Bhanushali

AI Engineer | LLM Applications · RAG · Model Fine-Tuning · Vector Databases · Machine Learning · Deep Learning

408-592-6951 | aryanbappsjobs@gmail.com | Aurora, IL | LinkedIn | GitHub | Portfolio

Summary

AI/ML Engineer bridging applied GenAI and core deep learning. Ships end-to-end LLM systems, fine-tuning (LoRA/QLoRA), LangGraph and MCP-based agents, retrieval, and evaluation, and builds neural architectures from scratch in PyTorch across CNNs, attention, multimodal transformers, and deep RL (DQN/PPO), with a peer-reviewed CNN publication. MS in Artificial Intelligence at Northeastern.

Experience

UST *Aliso Viejo, CA (Remote)*

AI Engineer Intern *Jul. 2026 - Present*

- Rebuilt a broken epic/story generation agent on an inherited multi-agent product-management platform into a **two-phase**, adaptive **LLM** pipeline (**Python**, **LangGraph**, **AWS Bedrock**, **Pydantic**) that decomposes an idea into an epic then authors each child story in its own call, cutting a product owner's manual epic-and-story drafting from **hours** to **under 10 minutes**.
- Designed a **four-scope** context system (Personal, Project, Workflow, Agent) with multimodal ingestion (text, image, **PDF**, voice) that folds the right context into each agent node at run time, so product owners can ground any agent in the exact source material a task requires.
- Built a guardrailed, scope-locked platform assistant (prompt-injection defenses, per-user isolation) and a **workflow composer** that turns a plain-English request into a valid, tier-ordered agent graph, letting non-technical product owners assemble and run pipelines without wiring nodes by hand.
- Built and tested a **Jira Cloud** integration that pushes AI-generated epics, stories, and story-maps into Jira as native issues, using an **Atlassian Document Format** serializer, per-user **pgcrypto**-encrypted credentials, and idempotent create-vs-update so re-pushing updates issues instead of duplicating them.
- Extended the platform's agent ontology, a **four-tier** model with tier-compatibility rules governing which agent can feed which, built real validator agents for every quality-tier output, consolidated **11 strategy agents** behind a shared runner, and rewrote **8** of them to emit typed structured output (quadrant matrices, issue trees, flow graphs) rendered natively in-app instead of prose.
- Added run-output persistence (**JSONB**) and a **JWT**-protected API so external coding agents can fetch generated specs, and contributed to the platform's Dockerized **CI/CD** (**GitHub Actions**, **Kubernetes**, migration to **GitHub OIDC**) alongside DevOps.

FounderWay.ai *Cambridge, MA (Remote)*

AI Engineer Intern *Jun. 2025 - Aug. 2025*

- Built an event-driven **agent orchestration** system as the sole AI engineer, using a **GPT-4o** classifier to route incoming **Vapi** call transcripts through filter rules to specialized worker agents over **HTTP** webhooks, turning raw transcripts into structured, task-routed work automatically.
- Engineered the **GPT-4o** document-generation worker serving **30+** users, grounding each document in the caller's context via multi-table **PostgreSQL** joins on user ID, cutting post-call report turnaround from ~**15 minutes** to **3 to 5** (an estimated ~**80%** reduction).
- Added output validation and fault tolerance using a router node that retried generation once on **JSON**-schema mismatch before returning a graceful error, and integrated **Gmail** to deliver completed documents as **PDF**, hardening the flow against malformed model output.

Projects

SportsBrain: MCP-Powered Football Scouting Agent | *MCP, LangGraph, RAGAS, Redis, AWS* *Feb. 2026 - Jul. 2026*

- Authored **3 MCP servers** (**FastMCP**) exposing **9 tools** over **2,839** FBref players, **508** Transfermarkt valuations, and Statsbomb match events in **SQLite**, and wired a **LangGraph ReAct** agent (**Qwen3 8B** via **Ollama**) to consume them over stdio through a **MultiServerMCPClient**, enabling plain-English scouting queries that span all three data sources over the **MCP** protocol.
- Evaluated the agent on a **163-question** golden set using **RAGAS** (**0.83** answer relevancy, **0.66** faithfulness, local **Qwen3 8B** judge) with every run traced in **LangSmith**, and cached tool results in **Redis** (**1h**/**24h** TTLs), achieving **162 of 163** error-free completion with sub-millisecond cached lookups.
- Containerized the platform with **Docker**, pushed the image to **Amazon ECR**, and deployed it to **AWS EKS** (**Kubernetes** Deployments, a **LoadBalancer** Service, ConfigMap, Secret), validating health, UI, and cache connectivity on the cluster, and authored a **GitHub Actions** build-to-ECR-to-EKS pipeline to prove the cloud path end to end.

Netflix Semantic Recommendation Engine | *Llama 3.2, QLoRA, ChromaDB, LangGraph, FastAPI* *Feb. 2026 - Jul. 2026*

- Built a two-stage semantic movie recommender using base **sentence-transformer** embeddings to retrieve candidates from **4,391** movies in **ChromaDB** and a fine-tuned **cross-encoder** to re-rank them, served through a **4-node LangGraph** agent (query parsing, vector search, re-ranking, explanation) in a Dockerized **FastAPI** app, delivering ranked, explainable recommendations.
- Fine-tuned **Llama-3.2-3B** as a **QLoRA** cross-encoder re-ranker (**4-bit NF4**, **r=16**, sequence-classification head) on ~**42K** query-candidate pairs with cosine-mined hard negatives, improving **NDCG@10** from **0.052** to **0.204** (~**4x**) over base retrieval on a held-out eval.
- Benchmarked cold-start performance against **SVD** collaborative filtering across **196** simulated new users, showing CF collapses to the same **10** popular titles while the semantic system personalizes, producing **782** distinct recommendations (a coverage and personalization win).

Spotify Emerging Artist Trend Detector | *Kafka, MongoDB, LangGraph, Ollama, Docker, FastAPI* *Feb. 2026 - Jun. 2026*

- Built a real-time A&R intelligence pipeline ingesting ~**490K** daily streaming events through **Apache Kafka** (**KRaft** mode) into **MongoDB** via an idempotent consumer-group upsert, then engineered a streaming-acceleration signal (ratio of **7-day** to **30-day** rolling velocity, normalized per artist) to surface artists whose momentum is accelerating.
- Built a **LangGraph ReAct** agent on **Llama 3.1 8B** (via **Ollama**) with **four** genre-aware **MongoDB** retrieval tools, letting A&R scouts query emerging-artist trends in natural language through a containerized **FastAPI** dashboard.

- Validated the acceleration signal against **4 years** of Spotify US Viral 50 chart data, flagging artists ahead of their chart debut with a median lead of **2 days** (up to **500+** at the extreme).

Semantic Turn Detection for Voice AI | *Llama 3.3 70B, Groq, Silero VAD, FastAPI, WebSockets* *Jan. 2026 - Apr. 2026*

- Built the semantic scoring engine for a voice-agent turn-detection system that decides when the AI should respond based on whether the speaker has finished their thought, replacing fixed silence-threshold **VAD** with an **LLM** ensemble that judges utterance completeness rather than mere silence.
- Engineered a **three-signal ensemble** on **Groq**-hosted **Llama-3.3-70B** (a syntactic end-of-utterance scorer, an intent classifier driving dynamic silence thresholds, and a context-aware completion scorer) combined into a weighted fire decision, paired with **Silero VAD** and **Groq Whisper**.
- Validated the ensemble in a within-subjects user study (**n=18**, 7-point Likert) against a silence-only **VAD** baseline, raising perceived naturalness from **4.31** to **6.00** and cutting user frustration from **3.08** to **1.53**.

Hierarchical Attention Pyramid (HAP) | *PyTorch, ResNet-50, ViT/DeiT; single-author working paper* *Dec. 2025 - Present*

- Designed a novel multi-scale attention module (three cascaded local-attention blocks plus a parallel global channel-attention branch, fused via adaptive pooling and **1x1** convolution) as a modular, drop-in replacement for the **ResNet-50** classification head, reusable on any CNN backbone.
- Benchmarked **ResNet-50+HAP** against **ResNet-50**, **ViT-Small**, and **DeiT-Small** on **CIFAR-10/100** across four data fractions, reaching **97.11%** (CIFAR-10) and **83.94%** (CIFAR-100) with only **ImageNet-1K** pretraining (**1.3M** images), competitive with transformers needing **ImageNet-21K** (**14M** images).
- Backed the result with data-efficiency analysis, parameter and training-time comparisons, and **Grad-CAM** visualizations confirming HAP learns complementary local and global features; sole author of the working paper.

HART: Hybrid-Input Revenue Transformer | *PyTorch, Transformers, Multimodal Fusion* *Nov. 2025 - Dec. 2025*

- Designed the architecture for a multimodal event-revenue forecasting transformer: modality-specific encoders (structured features, a **Transformer** time-series encoder, engineered signal proxies), an **MLP-Mixer** fusion block, and **cross-modal attention** with learnable per-modality gating.
- Caught fundamental data-integrity failures in the source dataset (recorded attendance exceeding stadium capacity, and target leakage through a revenue-derived series) and flagged the resulting metrics as invalid rather than reporting an inflated score.

Deep RL: DQN, Double DQN & PPO from Scratch | *PyTorch, Atari, HPC* *Feb. 2026 - Apr. 2026*

- Implemented **DQN**, **Double DQN**, and **PPO** from scratch in **PyTorch** (no RL libraries) with full Atari preprocessing, experience replay, and target networks, training ~**3M** frames per configuration across **12 runs** (3 algorithms x 2 seeds x 2 games) on the Northeastern **HPC** cluster.
- Ran a comparative study of value-based vs. policy-gradient methods on **Breakout** and **Enduro**, including a Q-value overestimation analysis contrasting **DQN** and **Double DQN** to characterize where each performs best.

SIGNN: Real-Time Gesture Recognition | *PyTorch, OpenCV, Streamlit, HAGRID-V2* *Oct. 2025 - Nov. 2025*

- Built a **CNN**-based real-time hand-gesture recognition system, lifting test accuracy from **47%** to **78%** by moving from a 3-block baseline to a deeper **BatchNorm** architecture with task-specific data augmentation (rotation, flip, color jitter).
- Deployed a real-time **Streamlit** webcam app mapping recognized gestures to computer controls, a touchless HCI/accessibility interface demonstrating end-to-end training through deployment.

Scientific Article Reference Predictor | *T5, Hugging Face, Multi-XScience* *Jan. 2025 - Apr. 2025*

- Fine-tuned **T5** (sequence-to-sequence) on the **Multi-XScience** corpus to generate the likely reference set for a scientific article from its abstract and related-work context, building the full **NLP** pipeline (tokenization, sequence formatting, training loop, learning-rate finder).
- Explored **BERT** tokenization and evaluation tradeoffs for the task, demonstrating task-adapted fine-tuning of pretrained transformers.

BI Dashboard | *Python, Gradio, pandas, Plotly* *Nov. 2025 - Dec. 2025*

- Built and deployed an interactive business-intelligence dashboard (**Python**, **Gradio**) that lets non-technical users upload CSV/Excel data and explore it through summary statistics, correlation analysis, interactive filtering, and multiple visualization types, hosted on **Hugging Face Spaces**.
- Implemented the automated-insights layer, top-N group ranking by configurable aggregation and **z-score**-based outlier detection, with defensive handling for empty files, missing columns, and non-numeric inputs.

Image Mosaic Generator | *Python, NumPy, Pillow* *Sep. 2025 - Nov. 2025*

- Built a photo-mosaic generator that reconstructs a target image from a tile library by dividing the image into a grid, computing per-cell average colors, and quantizing them to color buckets to select matching tiles.
- Engineered the pipeline to be fully vectorized in **NumPy** (grid averaging via array reshaping, tile assembly via fancy indexing, no per-pixel loops), keeping generation fast at high grid resolutions.

Skills

Languages: Python, SQL

LLM & Agentic AI: Large Language Models (LLMs), LangGraph, LangChain, Multi-Agent Orchestration, RAG, LLM Fine-Tuning (LoRA/QLoRA), Prompt Engineering, ReAct Agents, Model Context Protocol (MCP), LLM Evaluation (RAGAS), LangSmith, Hugging Face (Transformers, PEFT, bitsandbytes), Ollama, Groq

Machine Learning & Deep Learning: PyTorch, Transformers, Attention Mechanisms, CNNs, Reinforcement Learning (value-based & policy-gradient methods), Computer Vision, NLP (Natural Language Processing), Multimodal Fusion, sentence-transformers, scikit-learn, TensorFlow (familiar)

Data & Backend: FastAPI, REST APIs, Pydantic, Apache Kafka, Data Pipelines, PostgreSQL, MongoDB, Redis, SQLite, Vector Databases (ChromaDB, Pinecone), WebSockets, pandas, NumPy

Tools & Infrastructure: Docker, Docker Compose, Kubernetes, AWS (Bedrock, EKS, ECR), CI/CD (GitHub Actions), Git, Streamlit, Gradio, OpenCV, Make

Certifications: AI Fluency: Framework & Foundations (Anthropic, Jun. 2026); Claude 101 & Claude Code 101 (Anthropic, Apr. 2026); BCGX GenAI Job Simulation (Forage, Nov. 2025)

Publications

Convolutional Supernovae Networks | *JETIR, Vol. 11, Issue 4 (peer-reviewed)* *Apr. 2024*

- Co-authored a study benchmarking ML models for supernova classification from **PLAsTiCC** light curves, using Gaussian-process light-curve modeling and wavelet-based feature extraction; the best **CNN** reached **0.8425 AUC** with **0.99** recall.

Leadership & Involvement

Cogniscience (AI & Data Science Club) | *Co-Founding Member* *Jan. 2022 - May 2024*

- Co-founded the AI & Data Science department's student club, building club operations, technical programming, and student engagement from the ground up; ran ML/AI workshops for **30+** students and mentored **100+** peers on model development and technical career prep.

Climate Resiliency Hack | *2nd Place, Northeastern* *Dec. 2024*

- Placed **2nd** among **350+** participants across **7** Northeastern campuses in a climate-resiliency hackathon.

Education

Northeastern University *Boston, MA*

MS in Artificial Intelligence GPA 3.6/4.0 *Sep. 2024 - May 2026*

- Relevant Coursework: Deep Learning, Natural Language Processing, Reinforcement Learning, Machine Learning.

University of Mumbai *Mumbai, India*

BE in Artificial Intelligence & Data Science GPA 8.2/10 *Dec. 2020 - May 2024*