

Arvind Narayan

Staff AI/ML Engineer

hi@arwwwind.com · +91 91484 78347 · arwwwind.com · linkedin.com/in/arwwwind · github.com/arwwwind

ABOUT

Ten years building production ML and AI systems end-to-end — from data curation and model training to real-time inference on Kubernetes and AWS. Recently: multi-agent LLM platforms for scientific research at GATC Health (hybrid RAG, long-context memory, domain agentic tools). Before that: underwriting, pricing, and retail ML at Egen.ai (DriveTime, Carvana, Ahold Delhaize USA), education platforms at upGrad, and advertising systems at Yahoo. Depth in retrieval architectures, multi-agent workflows, LLM serving, graph neural networks, and clinical analytics — paired with distributed systems, microservices, and MLOps. Best in environments where AI has to leave the prototype and survive real users, real scale, and real impact.

TECHNICAL SKILLS

AI / LLM Systems	Multi-agent workflows (LangGraph), hybrid & Graph RAG, LIGHT-style memory, MCP, LangChain / LangSmith / Langfuse, LiteLLM, SGLang, fine-tuning & evals, guardrails, prompt engineering
ML / Deep Learning	PyTorch, GNNs (PyG), XGBoost / LightGBM, scikit-learn, Optuna, SHAP, feature engineering, SMOTE, ensembling, clinical ML
MLOps & Serving	Experiment tracking, model registry, CI/CD for ML, monitoring & drift, canary rollouts, SageMaker, AWS Bedrock, GPU inference, Qwen / Llama
Data Engineering	Databricks, Delta Lake / Lakehouse, AWS Glue, PySpark, Airflow, dbt, Kafka, Redshift, S3
Infrastructure	Kubernetes, Helm, Terraform, Docker (CUDA), Prometheus, Grafana, ELK, AWS
Languages & Backend	Python, TypeScript, JavaScript, Golang, FastAPI, Node.js, GraphQL, REST, microservices
Data Stores	PostgreSQL, Redis, MongoDB, Pinecone, pgvector, Weaviate, Chroma, OpenSearch
Application	React, TypeScript, Node.js, React Native, Tailwind, shadcn/ui

EXPERIENCE

GATC Health

Apr 2025 – Present

Senior AI/ML Engineer

Multi-Agent Workflows · Internal AI Research Platform

- Architected and shipped a production multi-agent AI research platform — an enterprise Perplexity for scientific teams — that became the primary AI interface for lab workflows within weeks of launch.
- Designed a hybrid sparse–dense retrieval pipeline (BM25/TF-IDF + dense embeddings, late fusion, cross-encoder reranking) for high recall on chemical and biological literature.
- Built a LIGHT-style memory subsystem that scales to millions of tokens of conversational history via episodic retrieval, structured working memory, and a compressed scratchpad — so multi-session research stays context-aware.
- Shipped domain agentic tools (toxicity lookup, molecular property prediction, structure normalization, ChEMBL & PubChem connectors) so agents can invoke specialized models and databases as callable actions.
- Served open-source models on SageMaker with SGLang; routed others through LiteLLM and AWS Bedrock for drop-in upgrades. Added evaluation harnesses and domain fine-tunes for alignment.

PAL-3 Post-Hoc Clinical Analytics · VistaGen Therapeutics

- Investigated a failed Phase 3 social anxiety trial by integrating longitudinal clinical outcomes, site operations data, enrollment timing, and turn-level speech features from recorded patient visits.
- Found that subject speech dynamics during treatment visits — conversation share, utterance length, total talk time — carried a consistent, leakage-aware predictive signal for placebo response; developed a habituation hypothesis explaining unexpected placebo performance.
- Explored recruitment channel effects, site-level calendar drift, psychometric symptom structure, and clinician vs. patient vocal dynamics as outcome moderators — with concrete recommendations for future trial design.

Graph Neural Networks for Molecular Property Prediction

- Designed and trained GNN architectures for toxicity, ADMET, and blood–brain barrier permeability prediction, choosing graphs over tree baselines to capture molecular topology — F1 ≈ 0.90, AUROC ≈ 0.92.
- Led end-to-end data curation from ChEMBL and PubChem: cleaning, deduplication, class-imbalance handling (SMOTE, weighted loss), and leakage-aware EDA.
- Deployed multi-modal ML models as low-latency GPU inference services on Kubernetes, consumed internally via documented APIs.

Egen.ai

Sept 2021 – Apr 2025

Senior ML Engineer

Financial Risk & Pricing ML · DriveTime · Carvana

- Led production ML for financial risk and pricing products used by DriveTime and Carvana — models that directly influenced underwriting and pricing across large origination volumes.

- Built predictive models for risk-adjusted APR, underwriting, LTV, depreciation curves, and delinquency — F1 scores near 0.90 across use cases, validated with rigorous backtesting.
- Owned customer data pipelines and feature engineering with client data teams: EDA, labeling strategy, feature importance, and fairness/bias checks for regulatory exposure.

Retail ML • Ahold Delhaize USA

- Delivered production ML for Ahold Delhaize USA and its retail brands (Food Lion, Giant, Stop & Shop, Hannaford, and others) — demand, pricing, and customer analytics across large multi-banner grocery operations.
- Partnered with client data and merchandising teams on feature pipelines, model validation, and rollout so recommendations stayed usable at store and banner scale.

Data Engineering & ML Inference Platform

- Built batch pipelines and analytics warehouses feeding underwriting and retail models — S3 landing, AWS Glue / PySpark transforms, curated features in Redshift.
- Led a scalable microservices inference platform on AWS Kubernetes (Helm + Terraform), with ELK observability, CUDA Docker images, and auto-scaling for GPU inference.
- Managed and mentored engineers; drove MLOps defaults for experiment tracking, model registry, monitoring, and canary releases.

upGrad

Dec 2019 – Sept 2021

Lead Software Engineer

LMS Rebuild • Full-Stack Platform

- Led the full-stack rebuild of a Learning Management System serving 3M+ active learners — ~75% improvement in Core Web Vitals via architecture, code splitting, caching, and PWA / offline-first delivery.
- Shipped React + TypeScript learner UX with Node.js and Golang services; pushed core API responses below 200ms with edge-aware CDN and Redis caching.

Assessments & Quiz Platform

- Built an online test-taking and quiz-authoring platform for large-scale peak load — 50–60k concurrent test-takers at peak hour, while off-peak traffic was mostly teachers authoring quizzes.
- Teacher side: question bank with search, test creation and invitation flows, plus a live demo mode to preview how assessments work end-to-end.
- Student side: quiz mode with proctoring controls — tab-switch detection, webcam-based facial distraction checks, and copy-paste restrictions — for secure high-stakes exams.

Shorts • Personalized Micro-Learning

- Conceived and built Shorts end-to-end — a TikTok-style micro-learning feed combining SM-2 spaced repetition with a feed-forward network for personalized sequencing and retention scheduling.
- Developed models that flag students likely to fail or drop out months ahead from engagement, attendance, and social signals — enabling proactive advisor intervention.

Yahoo

Mar 2018 – Dec 2019

Software Engineer

Yahoo Ad.com • Self-Serve SMB Advertising

- Contributed to Yahoo Ad.com, a self-serve platform for SMBs to create and manage campaigns across the Yahoo network with minimal setup.
- Built ML-assisted onboarding that used website URL, business name, and category to auto-generate campaign configs — reducing time-to-first-campaign.
- Built a business knowledge graph from websites, metadata, and third-party sources to auto-select CTA, audience, geo, and bidding — improving activation for first-time advertisers.

Yahoo Ad.com Mobile • React Native (ex-Yahoo Gemini)

- Shipped the Ad.com mobile app for iOS and Android in React Native — the same self-serve advertising product as Ad.com on the web, previously launched as Yahoo Gemini for mobile.
- Delivered campaign creation and management flows on mobile so SMB advertisers could run the full Ad.com workflow from their phones.

EDUCATION

Bachelor of Engineering, Computer Science — VTU — 2018

RECOGNITION

Star of upGrad — 2021 — Awarded for outstanding technical leadership and platform impact

upGrad Hackathon Winner — 2020 — Built the winning prototype for the Shorts micro-learning product