Vaishak Menon

La Jolla, San Diego, CA | +1 858-220-6008 | mvaishak.works@gmail.com | LinkedIn | GitHub

EDUCATION

University of California – San Diego **Sep 2025 – Jun 2027**
Master of Science, Data Science GPA: 4.0 / 4.0 San Diego, CA

- Coursework: Machine Learning, Statistical Modeling, Web Mining & Recommender Systems, Efficient AI, ML Systems

Vellore Institute of Technology **Jan 2017 – Jun 2021**
Bachelor of Technology, Electronics & Communication Engineering GPA: 3.79 / 4.0 Chennai, India

- Coursework: Neural Networks & Fuzzy Control, Linear Algebra, Probability & Random Processes

EXPERIENCE

Tredence Inc. | *Machine Learning Engineer* **Jan 2023 – Jun 2025**

- Deployed **Neural Collaborative Filtering** and a **contrastive learning** discovery engine to production, driving an **18% lift in recommendation CTR** and **14% improvement in add-to-cart rate** on the e-commerce platform.
- Built **customer scoring models** using **XGBoost** and **LightGBM** on large-scale behavioral event data to predict purchase propensity and churn risk; outputs were integrated into CRM systems to trigger retention workflows, improving retention by **11%**.
- Rebuilt model training from single-node to **distributed PySpark** pipelines on **Databricks**, enabling production-grade modeling on datasets exceeding **500M rows** and reducing training time by **85%**.
- Designed **A/B testing** infrastructure with **K-Means** user stratification to evaluate model changes in production; achieved a **6% NDCG@10 lift** and **8% recall gain** across live traffic, communicating results to product and engineering partners.
- Designed and deployed **MLWorks**, a production **model observability** platform for real-time **data drift**, bias, and quality monitoring; reduced ML issue detection latency by **60%**.

Tredence Inc. | *Data Analyst* **Jun 2021 – Dec 2022**

- Built and deployed a production **demand forecasting** platform using **time-series** ensembles (**FBProphet**, **ARIMA**, **SARIMA**), achieving a **10% MAPE improvement** and **8% inventory cost reduction** across global markets.
- Developed **customer segmentation** models using **RFM analysis** and **K-Means** clustering on transactional data, enabling targeted campaigns that improved conversion rates by **12%** across high-value cohorts.
- Automated global supply chain **KPI reporting** with **Power BI** dashboards, reducing executive reporting turnaround by **25%**.

Tredence Inc. | *Data Analyst Intern* **Jan 2021 – May 2021**

- Built a **Multi-Touch Attribution** model using **Markov Chains** to quantify digital channel contributions, improving **marketing budget efficiency** by **15%** across digital campaigns.

SKILLS

Programming: Python, SQL, PySpark, R, C++
Libraries & Frameworks: PyTorch, TensorFlow, Scikit-learn, XGBoost, LightGBM, Statsmodels, Hugging Face Transformers, LangChain, LangGraph, Pandas, NumPy, Plotly
DS/ML/DL: Statistical Modeling, Deep Learning, Transformers, Neural Collaborative Filtering, Contrastive Learning, Recommender Systems, Clustering, Time-Series Forecasting, A/B Testing, Feature Engineering, Embedding Models, NLP, Diffusion Models (DiT)
GenAI/LLM: LLMs, RAG, Multi-Agent Systems, Agentic AI, Prompt Engineering, Knowledge Graphs (GraphRAG), KV-Cache Optimization, Vector Databases (Qdrant), Semantic Search
MLOps/Cloud: MLflow, Databricks, AWS, Docker, Git, CI/CD, Langfuse, DeepEval, RAGAS, Experiment Tracking, Model Versioning, Model Drift Monitoring, Real-Time Observability

PROJECTS

PlayNext – Steam Game Recommender System | *github.com/mvaishak/PlayNext*

- Built a **hybrid collaborative filtering** recommender on the Steam dataset (**88,310 users, 32,135 games, 5.8M+ interactions**) using **bundle co-occurrence** as an auxiliary signal alongside user–item interaction history; achieved a **79.5% Hit Rate@10**, outperforming popularity baselines by **21 pp**.
- Evaluated across **Precision@K**, **Recall@K**, and **NDCG** on an 80/20 per-user split over **62,936 test users**; tuned hybrid score blending via grid search and deployed an interactive **Streamlit** demo with precomputed recommendations.

Scientific Literature Intelligence Platform

- Built a **multi-agent RAG** system with **LangGraph** orchestration, 2-hop **citation-graph traversal** (NetworkX), and **GPT-4o** synthesis with per-claim source attribution; demonstrated $\geq$**10 pp citation recall improvement** over vector-only RAG on a 50-question benchmark evaluated with **RAGAS** and **DeepEval**.
- Engineered a **LaTeX-aware chunking** pipeline for equation-preserving ingestion, **GPT-4o-mini** entity annotation, **Qdrant** vector store with **Cohere** reranking, and **Langfuse** observability traces logging token cost and retrieval hops per query.

KV-Cache Quantization for Long-Horizon Video Generation

- Conducted a 33-method empirical study of **KV-cache compression** and **quantization** for self-forcing video generation (**Wan2.1**), benchmarking peak VRAM, runtime, compression ratio, and **VBench** fidelity across methods.
- Identified **FlowCache-inspired soft-prune INT4** as optimal, achieving **5.4× compression** and reducing peak VRAM from **19.3 GB to 11.7 GB** while preserving generation quality in long-horizon inference.