AVIRAL UPADHYAY

Sunnyvale, CA 94085| (213) 608-7284| aviralupadhyay18.au@gmail.com| linkedin.com/in/aviralupadhyay | github.com/avral1810

PROFESSIONAL EXPERIENCE

Software Engineer, Machine Learning (Ads), Meta, Menlo Park, CA Jan 2024-Present

- **Architected** sub-millisecond early-stage retrieval models ranking 10M+ ads by building inverted-index–derived user–ad interaction features with sparse overlap signals and precomputed in-memory lookups.
- **Built** analysis and debugging workflows to diagnose failures caused by representation–loss mismatch on highly skewed targets.
- **Designed** and trained multi-objective ranking models under delayed and noisy feedback, optimizing long-horizon outcomes (purchase vs post-install engagement).
- **Developed** loss formulations enabling training on partially observed labels, reducing effective label latency from 8 days to 2 days while preserving gradient quality and offline–online alignment.
- **Built** training setups and evaluation workflows that simulate long-horizon decision making under sparse and delayed rewards.
- **Led** and **mentored** a small team (2–3 engineers) through model design, experimentation, and deployment.

Software Development Engineer, Amazon Ads, New York, NY Dec 2022-Jan 2024

- **Built** and optimized data and inference pipelines for Falcon-7B LLMs, focusing on representation, grounding, latency, and failure-mode debugging in production systems.
- **Designed** large-scale NLP data processing pipelines to generate training and inference representations from raw web data, reducing operational cost by ~60%.
- **Developed** a CDK based IaaS pipeline to provision a scalable distributed compute infrastructure in AWS (**TypeScript**)
- **Decreased** operational cost of the Amazon's internal bidding system by 40% by using a **Bloom Filter** to clean out low performing webpages (**Java**)
- **Designed** a CICD pipeline for deployment of Data Pipeline on AWS (**Typescript**)

Software Engineer, TikTok/ByteDance, Mountain View, CA Jun 2021-Dec 2022

- **Built systems** to extract salient frames from live streams for efficient review and moderation.
- **Developed** statistical forecasting models (ARIMA) for traffic prediction and capacity planning.
- **Designed** distributed scheduling and throttling systems to reduce upload failures from ~1000/day to near zero.
- **Built** real-time infrastructure in Go supporting CDN scheduling and global live streaming.

Data Solution Engineer Intern, Wayfair LLC, Boston, MA (working remotely from Los Angeles, CA) Jan 2021–May 2021

- **Implement** ETL pipelines to report customers' behavior on Wayfair's website and migrate it from Hive to Big Query (**Big Query, Python, HIVE**)
- **Examined** user data to provide insight on ad activity on Wayfair's website and ad platforms to the Marketing team (**Spark**, **PySpark**)

EDUCATION

University of Southern California, Los Angeles Aug 2019–May 2021
Master of Science, Computer Science **CGPA: 3.66 / 4**
Coursework: Foundations of Artificial Intelligence, Analysis of Algorithms, Machine Learning for Data Science, Applied Natural Language Processing, Applied Machine Learning for Games, Information Retrieval and Web Search Engines, Database Systems

Vellore Institute of Technology, Tamil Nadu, India Jul 2014–May 2018
Bachelor of Technology in Computer Science **CGPA**: **8.90 / 10**
Coursework: Data structures and Algorithms, Agent-Based Intelligent Systems, Cloud Computing, Operating Systems, Embedded System

TECHNICAL SKILLS

Languages: Python, Go, C/C++, Java **Infra:** AWS (S3, EMR, EC2, Step Functions), Linux **Data:** Spark, PySpark, Presto, SQL, BigQuery, **ML:** PyTorch, deep learning, regression, multi-task learning, ranking systems, LLM inference

PROJECTS

Vision-Based RL Agent (FIFA)

- **Trained** an agent from raw pixels using CNN + LSTM without access to environment state or game engine internals.
- Learned control policies from sparse and delayed rewards through trial-and-error interaction with the environment.

Stick Man Human (Python/C++) Jan 2021-May 2021

- **Built** OpenGL-based 3D human model driven by motion capture.
- **Used** MediaPipe to extract joint coordinates and map human motion to 3D model.

Human Gait analysis and Parkinson's Disease (Python) Jan 2018-May 2018

- **Used** sensor data and Random Forest to classify Parkinson's disease.
- **Built** neural and YOLO-based baselines for gait analysis.

Software Engineer

AVIRAL UPADHYAY

Sunnyvale, CA 94085| (213) 608-7284| aviralupadhyay18.au@gmail.com| https://www.linkedin.com/in/aviralupadhyay

PROFESSIONAL EXPERIENCE

Software Engineer, Machine Learning (Ads), Meta, Menlo Park, CA Jan 2024-Present

- **Built** systems to predict and optimize advertiser and user value, directly impacting conversion and engagement metrics.
- **Architected** sub-millisecond early-stage retrieval models ranking 10M+ ads by building inverted-index–derived user–ad interaction features with sparse overlap signals and precomputed in-memory lookups.
- **Built** analysis and debugging workflows to diagnose failures caused by representation–loss mismatch on highly skewed targets.
- **Improved** ranking and delivery systems to increase user engagement and advertiser conversions through iterative experimentation.
- **Analyzed** experiment results and user behavior across cohorts and funnels to identify drop-offs, wins, and regressions, and used those insights to guide ranking and product changes.
- **Improved** feedback and attribution pipelines to make experiments reflect real user behavior faster.
- **Partnered** with product teams to design experiments that improved onboarding, engagement, and conversion for short-form video ads.
- **Led** and mentored a small team (2–3 engineers) shipping features and experiments that improved engagement and conversion.

Software Development Engineer (Ads), Amazon, New York, NY Dec 2022-Jan 2024

- **Built** and optimized data and inference pipelines for Falcon-7B LLMs, focusing on representation, grounding, latency, and failure-mode debugging in production systems.
- **Designed** large-scale NLP data processing pipelines to generate training and inference representations from raw web data, reducing operational cost by ~60%.
- **Developed** a CDK based IaaS pipeline to provision a scalable distributed compute infrastructure in AWS (**TypeScript**)
- **Decreased** operational cost of the Amazon's internal bidding system by 40% by using a **Bloom Filter** to clean out low performing webpages (**Java**)
- **Designed** a CICD pipeline for deployment of Data Pipeline on AWS (**Typescript**)

Software Engineer, TikTok/ByteDance, Mountain View, CA Jun 2021-Dec 2022

- **Built systems** to extract salient frames from live streams for efficient review and moderation.
- **Developed** statistical forecasting models (ARIMA) for traffic prediction and capacity planning.
- **Designed** distributed scheduling and throttling systems to reduce upload failures from ~1000/day to near zero.
- **Built** real-time infrastructure in Go supporting CDN scheduling and global live streaming.

Data Solution Engineer Intern, Wayfair LLC, Boston, MA (working remotely from Los Angeles, CA) Jan 2021–May 2021

- **Implement** ETL pipelines to report customers' behavior on Wayfair's website and migrate it from Hive to Big Query (**Big Query, Python, HIVE**)
- **Examined** user data to provide insight on ad activity on Wayfair's website and ad platforms to the Marketing team (**Spark**, **PySpark**)

EDUCATION

University of Southern California, Los Angeles Aug 2019–May 2021
Master of Science, Computer Science **CGPA: 3.66 / 4**
Coursework: Foundations of Artificial Intelligence, Analysis of Algorithms, Machine Learning for Data Science, Applied Natural Language Processing, Applied Machine Learning for Games, Information Retrieval and Web Search Engines, Database Systems

Vellore Institute of Technology, Tamil Nadu, India Jul 2014–May 2018
Bachelor of Technology in Computer Science **CGPA**: **8.90 / 10**
Coursework: Data structures and Algorithms, Agent-Based Intelligent Systems, Cloud Computing, Operating Systems, Embedded System

TECHNICAL SKILLS

Languages: Python, Go, C/C++, Java **Infra:** AWS (S3, EMR, EC2, Step Functions), Linux **Data:** Spark, PySpark, Presto, SQL, BigQuery, **ML:** PyTorch, deep learning, regression, multi-task learning, ranking systems, LLM inference **Experimentation**, A/B testing, metrics analysis, conversion optimization, attribution

PROJECTS

Vision-Based RL Agent (FIFA)

- **Trained** an agent from raw pixels using CNN + LSTM without access to environment state or game engine internals.