Arjun Rao

mailarjunrao@gmail.com | arjun.fyi | arjun-rao | arjunra0 | Google Scholar

Summary

Experienced software engineer with industry experience specializing in recommendation systems, and the intersection of applied research and ML infrastructure. My experience includes optimizing software systems for unlocking novel ML capabilities, improving researcher productivity, leadership experience with road mapping engineering initiatives, mentoring engineers, and hiring and academic research experience with peer-reviewed publications.

Industry Experience

Netflix

Los Gatos, CA, US

Software Engineer, Machine Learning

March. 2024 - Present

Building & optimizing training infrastructure for core ranking & page construction ML models, at the intersection of Applied Research and ML Engineering. Responsibilities include end-to-end ML training infrastructure design and implementation. As engineering partner for applied researchers, I collaborate with ML leadership, downstream serving & ML platform team engineers to unlock new capabilities, improve experimentation agility and optimize resource utilization. I'm also involved in the hiring process, conducting system design and technical interviews for various roles. Key Projects:

- Tech lead for migrating our existing training infrastructure to a novel managed ML experimentation framework that is being co-designed with the migration effort. Contributing to cross-cutting architecture design and system integration design to unlock capabilities, faster experimentation, cost-savings and improved compute & data sharing.
- Integrating a flexible dynamic catalog access system with our training infrastructure to enable access to accurate, point-in-time catalog availability at model training time to unlock improved personalization capabilities that support new content types and subscription tiers.
- Migrated our real-time signal events pipeline to a flexible framework that allows schema changes independent of binary deployments, mitigating the need to backfill data, leading to faster turn-around time for new user signal adoption from several weeks to few days.
- Prototyping a LLM based RAG solution for internal code documentation and Question Answering using Elastic Search, LangChain and Open-Source LLMs.
- Technologies Used: Java, Scala, Python, Tensorflow, Pytorch, S3, Iceberg, Spinnaker, Bazel, Monorepo

Microsoft Corp.

Redmond, US

Software Engineer II

June. 2021 - Feb 2024

Area lead for the online serving infrastructure for Microsoft's personalized feed product (MSN/Microsoft Start) that is surfaced across various canvases like Edge, Windows, Xbox and Skype (500M+ MAU). My role involved designing and spearheading various infrastructure initiatives that significantly improved system performance, capability, user engagement, and researcher productivity across the ranking stack. I was additionally involved with various initiatives to improve org culture, documentation discovery. I conducted several interviews, onboarded team members, was an intern mentor. Key Projects:

- Primarily owned the serving stack for a real-time feature store used for serving aggregated user & item features for ranking models. The service was implemented using an internal implementation of Flink, a high-throughput key-value store, and custom serving infrastructure. I was responsible for designing support for various aggregation capabilities including second-level sliding-windows, life-long time-decayed aggregation and long-term batch aggregation in collaboration with 2 other engineers. The serving system had an end-to-end avg. latency of less than 100ms (p95) and served an average 2M+ QPS from traffic across geographies. I was responsible for the end-to-end serving architecture & implementation, aimed at low-latency serving of fresh features, optimizing resource usage and supporting rapid experimentation agility. The system served 50% of most important features in the L2 Ranker, and contributed to 11% incremental revenue gains, aggregated 15% gain in engagement metrics. My direct impact included an aggregated 50% reduction in CPU & memory footprint over several projects across 2 years.
- Served as a tech lead for the debuggability workstream for FY24, working with a cross-functional team of engineers distributed across Egypt, China and Redmond. Worked on planning, delegation and identified gaps and strategies for reducing the time to debug issues across the recommendation stack. During the summer I was an intern mentor, where I planned our team's intern project, mentored the intern on a day-to-day basis and ensured a successful outcome.
- Architect for a new configuration compiler for Experiment Configuration Management that automated several manual steps required to setup & tear down A/B experiments and improved time from idea to config by several orders of magnitude (hours to few minutes). Played a principle role in designing the new workflow, SDK interfaces, and technical architecture for the overall project.
- Experimented with using GPT4 for User Interest Summarization and LLM based Feed Ranking as an internal prototype.
- Technologies Used: C++, C#, .Net, Python, Pytorch

Stride.ai Inc. *Bangalore, India*

NLP Engineer *Jul'18 - May '19*

- Designed and built re-usable end-to-end components to build custom Intelligent Process Automation solutions that leverage AI models & support document intensive processes for information extraction & compliance across industry verticals like financial regulation, banking and investments. Reduced the turn around time to build a new Proof-of-Concept from over 1 month to 1 week. The components I built included all parts of the stack from UI, Backend, Storage, Model Training, Inference, Feature Parsing and Serving.
- Built a custom PDF Viewer using PDF.js that allowed automatic ML based annotations and additional capabilities like support for multiple monitors, side-by-side scrolling of multiple documents, cross-browser support, lazy-loading documents. The viewer is was used in 10+ Projects and reduced latency of loading 100 page pdfs from 5+ seconds to < 1 second.
- Technologies Used: Python, Django, Keras, Tensorflow, D3.js, Angular, Typescript, PDF.js, Internal tools

Google Inc. *San Francisco, US*

Developer Programs Engineer Intern, Cloud DevRel *June 2017 - Aug. 2017*

Built internal tooling to help triage issues & bugs across open source repositories owned by Google Cloud. Part of the work was released as an open source project to demonstrate the use of various GCP offerings like Cloud Datastore, BigQuery, App Engine, Go and Angular.

Education

University of Colorado, Boulder *Boulder, US*

M.S. in Computer Science *Aug. 2019 - May 2021*

- Awarded Lloyd Botway Fellowship for Outstanding Master's students

Ramaiah Institute of Technology *Bangalore, India*

B.E. in Information Science and Engineering *Aug. 2014 - June 2018*

- **Best Outgoing Student** (Batch of 2014-2018), Dept. of Information Science & Engineering

Academic Research Experience

University of Colorado, Boulder *Boulder, US*

Emotive Computing Lab, Institute of Cognitive Science *Jan. 2020 - May 2021 (1.5 Years)*

- Worked with Prof. Sidney D'Mello as part of his Emotive Computing Group. Research focus was on discourse and language modeling for team-based collaborative problem solving.
- Developed a re-usable pytorch based framework for training and experimenting with transformer based language models and multi-modal temporal sequential learning models using LSTMs.
- Co-authored 5 Peer-reviewed publications submitted to top-tier conferences and journals.
- Technologies Used: **Python**, **PyTorch**, **GPU Training AWS**

Select Publications

CPSCoach: The Design and Implementation of Intelligent Collaborative Problem Solving Feedback *Tokyo, Japan / Virtual*

Angela EB Stewart, **Arjun Rao**, Amanda Michaels, Chen Sun, Nicholas D Duran, Valerie J Shute, Sidney K D'Mello *July. 2023*

- In Artificial Intelligence in Education AIED 2023 Lecture Notes in Computer Science [Link]
- We present the design of a fully-automated system that assesses and provides feedback on collaborative problem solving (CPS) competencies during remote collaborations

Do Speech-Based Collaboration Analytics Generalize Across Task Contexts? *Online, USA*

Samuel L Pugh, **Arjun Rao**, Angela EB Stewart, Sidney K D'Mello (Best Paper Nominee) *March. 2022*

- Proceedings of 12th International Learning Analytics and Knowledge Conference (LAK22), 2021 [PDF]
- We investigated the generalizability of language-based analytics models across two collaborative problem solving tasks, among 95 triads who collaborated over video conferencing.