

Balraj Ashwath

Machine Learning Engineer |  LinkedIn |  Google Scholar

 balrajashwath98@gmail.com

 +91-9964952869

EXPERIENCE

- ServiceNow** Bengaluru, IN
 - Senior ML Engineer** *Sep 2023 – Present*
 - MCP Server | Core Platform Frameworks Engineering (AI Control Tower)** *Jun 2025 – Present*
 - Core team member building ServiceNow's MCP Server from the ground up – exposing enterprise AI capabilities to MCP clients like Claude and ServiceNow's AI Agent Studio. Contributed end-to-end across the Python service, Kubernetes workload, and ServiceNow scoped application.
 - Reliability & Observability:** Founded the service's end-to-end telemetry, alerting, and on-call infrastructure (Prometheus, Grafana). Contributed to MCP-S release / deployment process & built core reliability primitives. Stood up isolated subprod infra with capacity planning to handle customer subprod-instance MCP-S traffic.
 - Core Feature Engineering:** Contributed across a wide range of MCP-S components – Redis caching, rate limiting, request throttling, circuit breakers, session management, JWT-based authentication, and service guardrails. Drove monthly releases, defect resolution, and performance improvements.
 - Platform Extensibility & Security:** Engineered customer MCP configurations to support Update Set migration and instance clones. Extended tool-creation framework to support REST APIs as MCP tools. Built file-attachment MCP tool for document upload to instance records. Spearheaded security hardening – minimized container image security findings, resolved high-severity vulnerabilities, and shipped FIPS-compliant builds for regulated markets.
 - Customer Adoption & Documentation:** Authored **20+** Knowledge Base articles and client integration guides, driving MCP-S adoption across customers and internal BUs through faster issue resolution and setup. Kickstarted internal docs for release, on-call & alerting workflows. Resolved **25%** of customer-reported cases for MCP-S team.
 - Core AI / Advanced Technology Group** *Sep 2023 – Jun 2025*
 - Core contributor across the LLM development lifecycle for ServiceNow's internal foundation models (NowLLM family) – synthetic data generation, SFT and DPO / RLHF alignment, and large-scale evaluation across model checkpoints, AI skills, and languages.
 - Alignment Data Generation:** Designed **15+** categories of complex reasoning and instruction-following datasets for SFT and DPO / RLHF training, using and contributing to a custom internal graph-based synthetic data generation pipeline. Published preference-pair datasets that fed directly into model training.
 - Model Evaluation Infrastructure:** Built end-to-end auto-evaluation pipelines combining embedding-based representative sampling (high statistical confidence) and an automated LLM-as-judge framework for structured error analysis – surfacing strengths and weaknesses per metric across Faithfulness, Completeness, Relevance, and Coherence. Onboarded **12+** product AI skills to the evaluation platform, expanding coverage across summarization, generation, search, and workflow domains.
 - Now Assist Summarization:** Built preprocessing pipelines that improved Now Assist summarization quality and reduced hallucinations across HR, ITSM, and CSM business units in production. Conducted systematic evaluation studies on real-world records to track quality improvements and regressions.
- Qualcomm** Bengaluru, IN
 - Senior ML Engineer** *(Org: Corporate R&D Qranium AI SW)* *Apr 2022 – Jun 2023*
 - Part of ML Accuracy Team for Qualcomm's AI Cloud division working to optimize ML models and tuning parameters to meet accuracy targets on Qualcomm Cloud AI 100 Inference Accelerator (AIC 100). Performed measurement, analysis and optimization of accuracy of various ML models for INT8 quantization to obtain the best Performance vs Accuracy trade-off and achieve INT8 accuracy targets of **99%** of FP32 accuracy baselines.
 - AI Model Efficiency Toolkit (AIMET):** Involved in the development of new quantization techniques. Implemented new learnable calibration schemas to enable higher accuracy gains from Quantization-aware fine-tuning (QAT). Conducted AIMET experiments related to Post-Training Quantization (PTQ) techniques such as CLE / BC, Adaptive Rounding and Auto Mixed Precision. Aided in evaluation of AIMET and AIMET Pro SW for different CV, NLP models. Supported enabling more QAT models inside AIMET zoo.
 - Migration of various models' mixed-precision configs originally developed for AIC 100 SW stack to support QNN (cross-device API for machine learning) as a part of QNN-AIC Convergence. End-to-end handling of development, performing model evaluations and handling any regression issues found in Daily Accuracy & Performance Digests.

- Microland Limited - Automation Team** Bengaluru, IN
Senior ML Engineer & Data Scientist [Assoc Tech Arch - SW Dev] Jul 2019 – Apr 2022
 - Metric Time Series Analysis & Modeling:** Developed **Forecasting & Anomaly Detection** systems on metric data collected on edge devices. For *forecasting*, models such as *ARIMA*, *Prophet*, *NeuralProphet* were trained with a high ensemble MAPE. AutoEncoders were used to perform Anomaly Detection with reconstruction *F1* at **81%**. High precision of forecasted-anomalies aided effective monitoring of edge environments.
 - Alert Clustering:** Developed NLP-based **text clustering** system aimed at grouping unlabeled alerts into meaningful topics. Trained models such as *K-Means*, *LDA (TFIDF)*, *doc2vec*, *BERTopic* & other transformer-based models. *Tagging pipeline* was built to allow Ops Engineers to tag clusters and *thereby tagging any incoming alert to its closest cluster*. Post-processing methods including cluster correlation, new alert-type identification were applied to accurately tag more than **90%** of incoming alerts thereby *eliminating the need for a labeled alert corpus*.
- Vision and AI Lab, Indian Institute of Science** Bengaluru, IN
Research Intern (Mentors: Dr. R. Venkatesh Babu & K. Ram Prabhakar) May 2019 – Jul 2019
 - HDR Image Deghosting:** Developed novel CNN model for *motion compensation* in High Dynamic Range image-fusion & -deghosting. The proposed model improved SOTA on Khalantari SIGGRAPH 2017 dataset by **1.2 dB** PSNR (Peak Signal-to-Noise Ratio) score with **60%** faster inference times. [Work Presented at ECCV 2020]
- Computer Vision & Deep Learning Lab, NIT-K** Surathkal, IN
Undergraduate Researcher (Mentors: Dr. Shyam Lal & Karuna K. Eerapu) Jul 2018 – Apr 2019
 - Road Extraction from Satellite Imagery:** Developed novel segmentation architecture to identify road pixels in aerial imagery based on concepts from U-Net, dense & residual networks. The proposed model improved SOTA on Massachusetts Roads dataset by **2%** with an *order-of-magnitude lesser* parameters. [IEEE Access Journal Paper]
- RBL Bank** Mumbai, IN
Summer Intern (Machine Learning) May 2018 – Jul 2018
 - e-KYC Process Automation:** Developed lightweight CNN-based OCR models to identify KYC details from PAN/ Aadhaar cards. Achieved **83%** F1 score with **sub-30ms** inference times on low-compute edge devices.

EDUCATION

- National Institute of Technology Karnataka** Surathkal, IN
B.Tech in Electronics & Communications Engineering | GPA: 7/10 | Senior Thesis: 10/10 Aug 2015 – Apr 2019
- Narayana Pre-University College** Bengaluru, IN
Karnataka State Pre-University | Percentage: 90% | JEE Mains: 99th %ile Jul 2013 – May 2015

PUBLICATIONS

- Dense Refinement Residual Network for Road Extraction from Aerial Imagery Data**, Karuna K. Eerapu*, Balraj Ashwath*, Shyam Lal, Fabio D. Acqua & A.V. Narasimhadhan, *IEEE Access Journal*, 2019. [* - equal contbⁿ]
- Towards Practical and Efficient High-Resolution HDR Deghosting with CNN**, K. Ram Prabhakar, Susmit Agrawal, Durgesh Singh, Balraj Ashwath, & R. Venkatesh Babu, *European Conference on Computer Vision*, 2020.

SKILLS

Languages: Python, JavaScript, SQL, Java, C++

AI / ML: PyTorch, HuggingFace, LLM Alignment (SFT, DPO, RLHF), LLM Evaluation, RAG, Prompt Engineering, Agentic AI, LangGraph, Scikit-learn

Infrastructure & Tools: FastAPI, Pydantic, Redis, Kubernetes, Prometheus, OpenTelemetry, Grafana, Splunk, Model Context Protocol (MCP), Git, Postman

HONORS & AWARDS

- Shortlisted in Microsoft's Incubathon 2021, Silver Medalist in Kaggle's NFL Impact Prediction 2021 - Top 10%
- Intel Edge AI Scholarship (2019), HackWithInfy Finalist (2018), Merit-cum-Means Scholarship, FFE Scholarship