Parth **Parmar**

parthparmar15@gmail.com | pparmar15.github.io/personal/ | parth-parmar

Summary

Software engineer focused on machine learning infrastructure and large-scale distributed systems. Experienced in building production systems for LLM-based products, including agent frameworks, synthetic data generation pipelines, and evaluation platforms for reinforcement learning and model benchmarking. Background includes cloud infrastructure and AI platform development at Amazon and in a startup environment.

Education

University of Toronto — *Toronto, ON*

Bachelor of Applied Science in Engineering — *Sep 2012 – Apr 2018*

Minor : Computer Science

Skills

Languages	Python, Java, C++, Kotlin, C#, C
ML & AI	Ray, vLLM, VERL, FastAPI
Cloud & Infra	AWS, Docker, Terraform, Kafka, Kubernetes, Grafana, Prometheus
Frameworks	gRPC, Spring Boot, React, Node,

Work Experience

Amazon — *New York, NY*

Software Development Engineer (Amazon Ads) — *Sep 2024 – Present*

- Played a key role in the beta launch of an advertiser-facing AI agent that analyzes campaign performance data and generates actionable insights, improving advertisers' ability to plan and optimize campaigns
- Led development of an Ads AI agent framework on AWS Bedrock AgentCore and Strands SDK enabling multi-turn conversational agents with streaming responses, tool orchestration, MCP integrations, and memory persistence supporting short-term and long-term context with per-session and per-advertiser isolation, allowing agents to maintain conversational state across interactions
- Architected and built a distributed synthetic data generation platform that produced millions of high-quality LLM training samples across thousands of advertisers, enabling faster post-training workflows (SFT, DPO) and overcoming cold-start constraints by programmatically producing multi-step ReAct reasoning traces with tool use via teacher models
- Developed a scalable LLM-as-Judge evaluation platform generating real-time reward signals for RL training and model evaluation pipelines, supporting high-throughput synchronous and asynchronous batch processing across thousands of evaluations

Aryn.ai — *Mountain View, CA*

Software Development Engineer (Founding Team) — *Jul 2023 – Aug 2024*

- Collaboratively developed an AI-powered open source document processing framework (Sycamore) for RAG and unstructured analytics using distributed Python framework Ray
- Led the development of a Minimum Viable Product (MVP) for an inference service using an in-house developed DETR model, enabling users to segment and partition their PDFs efficiently
- Spearheaded the design and development of the control plane for managed offering of the Sycamore platform enabling the company to attract early adopters
- Regularly reviewed code and provided mentorship to junior developers, fostering a culture of quality and continuous improvement within the engineering team

Amazon Web Services — *Palo Alto, CA*

Software Development Engineer (Lake Formation & Glue Elastic Views) — *Feb 2020 – June 2023*

- Successfully launched a beta version of a managed service, providing data storage optimization for consumer data lakes, resulting in improved data storage efficiency and cost savings
- Worked and collaborated in a cross-functional team environment to launch ACID transactions support in S3 data lakes
- Early team member of AWS Glue Elastic Views - a service that allows customers to use familiar SQL to easily create materialized views from multiple different data sources without the hassle of managing any infrastructure
- Led the design and development of on host TLS termination for two major services to achieve Encryption in transit
- Collaborated with senior software engineers to design and develop multi-tenancy Incremental-View-Maintenance Service using various AWS technologies

Flipp — *Toronto, ON*

Software Engineer Intern (Digital Content Production) — *Sep 2019 – Dec 2019*

- Built TypeScript, Node, and Kafka microservice features for a new digital publishing system, contributing to architecture and reducing flyer generation time from 12 hours to about 1 hour

University of Toronto — *Toronto, ON*

Research Intern (Computational Linguistics Lab) — *Apr 2018 – Aug 2018*

- Developed a Python-based NLP and machine learning pipeline to classify cause of death from symptom descriptions and evaluate algorithms for improved accuracy

Publications

- Co-authored the paper "The Design of an LLM-powered Unstructured Analytics System" (arXiv:2409.00847) during my time at Aryn for Conference on Innovative Data Systems Research (CIDR)