


PRAJWAL V. ATHREYA


+1(959)-237-0577 | athreya.p@northeastern.edu | <https://www.linkedin.com/in/prajwal-v-athreya> | Portfolio 
Available for full-time opportunities from May 2025

A graduate student specializing in artificial intelligence, with nearly two and a half years of industry experience. Skilled in **Python, Go, C++, and Java**; actively seeking full-time opportunities as a Software Development Engineer (SDE).



EDUCATION

Northeastern University, Boston, MA	Sep 2023 - May 2025
Master of Science in Artificial Intelligence	GPA: 3.9
Dayanandasagar College Of Engineering, Bangalore, KA	Aug 2017 - Aug 2021
Bachelor of Engineering in Automobile Engineering	GPA: 3.7

EXPERIENCE

Software Engineer (Co-op) - Akamai, Cambridge, MA 	July 2024 - December 2024
· CAPL (Cluster API Provider Linode)	
· Added E2E tests using Chainsaw for multiple flavors of kubernetes clusters; increasing test coverage by almost 20%	
· Implemented Linode Cloud Firewall and Placement groups as a default resource for clusters.	
· CSI Driver (Container Storage Interface)	
· Optimized GOMAXPROCS to dynamically scale resource utilization for improved performance during CSI driver operations.	
· Added support for Block Storage Encryption of volumes.	
· Integrated Prometheus and Grafana for visualizing and monitoring Linode Cloud Volumes related CSI driver operations.	
· Implemented OpenTelemetry to enable tracing of gRPC function calls and improve observability and stack tracing.	
Software Engineer - UpUgo & Surgg Pvt Ltd, Bangalore, KA	November 2021 - July 2023
· Workout Recommendation System	
· Engineered a hybrid recommendation system that combines collaborative filtering and transformer-based embeddings to enhance accuracy and personalization.	
· Improved recommendation system accuracy using LoRA ; which significantly improved retraining time.	
· Containerized the service using Docker and deployed it with Amazon EKS(Elastic Kubernetes Service) for scalable orchestration and seamless load balancing.	
· Integrated TensorRT into containerized inference pipelines, reducing the recommendation generation time from 470 to less than 200 ms , enhancing real-time performance.	
· Implemented in-memory caching and CDN-based caching using Cloudflare for APIs resulting in improved retrieval speeds of approximately 30% .	

PROJECTS

Adversarial Attacks on Large Language Models 	March 2024
Related Topics: Positional Encoding, Transformers, GPT, Model Fine-tuning, Generative Model, pEFT	
· Fine-tuned GPT-2 to generate adversarial inputs, successfully causing misclassification in 60% of cases .	
· Deployed this generative model with FastAPI as the backend framework to convert the inference module into an API, allowing it to be called upon text for testing adversarial examples.	
Image Processing Application 	October 2023
Related Topics: MVC Architecture, Multi-Threading, Multi-Processing, Object-Oriented Design	
· Built a full-scale image processing application employing MVC architecture in Java.	
· Following the said architecture and minor optimizations during pre-processing processing times were reduced by approximately 70ms, improving the performance by almost 40% .	

SKILLS

Programming Languages	Bash, C++, Java, Node.js, Python, Go
Database	MongoDB, MySQL
DevOps	Docker, EC2, EKS, Elastic Beanstalk, Git, Lambda, S3, Chainsaw
Operating Systems	Linux, Windows