Aditya Padgal

+1 (930) 333-4229 | adityapadgal@gmail.com | linkedin.com/in/aditya-padgal | github.com/adityapadgal

Education

Indiana University Bloomington — Bloomington, IN
MS in Computer Science — *Aug. 2023 – May 2025*

University of Pune — Pune, India
BE in Information Technology — *Aug. 2017 – April 2021*

Experience

Machine Learning Engineer — Sep. 2024 – Present
Indiana University Bloomington — *Bloomington, United States*

- Built an LLM-powered RAG clinical assistant with LangChain, PyTorch, and FAISS to retrieve 150K+ clinical trials, assess patient eligibility with DeepSeek-r1:8B, and handle 5K queries/day with structured JSON outputs.
- Optimized inference using P-Tuning, Thread-of-Thought, and LLM Finetuning, leveraging PyTorch for batching and optimized attention mechanisms, reducing hallucinations by 70% and improving response consistency.
- Developed NeuroBERT, a PubMedBERT-based NLP model trained on 1M+ clinical records (NACC, UK Biobank, ADNI) to extract insights from neuropathology reports, reducing manual review by 40% and accelerating Alzheimer's research.
- Accelerated text processing with TensorRT, CUDA, reducing inference time by 3x (2s to 500ms) via batch processing and multi-threading.

Data Engineer — Jan. 2021 – Aug. 2023
Quantiphi, Inc — *Mumbai, India*

- Designed a hybrid cloud disaster recovery solution with multi-region, active-active storage using Amazon S3 and Azure Blob Storage, implementing automated geo-replication to achieve 99.98% uptime.
- Built and optimized scalable data pipelines for batch and real-time processing using AWS Step Functions, Azure Data Factory, Snowflake, AWS Lambda, and PostgreSQL, leveraging Change Data Capture (CDC) with AWS DMS to reduce latency to sub-second levels, cut processing time by 73%, and save $500K annually.
- Developed a serverless data processing system using AWS Lambda, AWS DMS (CDC), and PostgreSQL, reducing data latency to sub-second levels while handling 400K+ events/sec.
- Optimized ETL process, reducing runtime from 5 hours to 30 minutes, and implemented an ML model using Amazon SageMaker improving data quality checks, cutting manual review time by 70%.
- Earned 4 awards and 6 nominations for leadership in cloud data engineering, automation, and real-time processing innovations by Quantiphi.

Software Developer Intern — July 2019 – Oct. 2019
Centre for Police Research — *Pune, India*

- Developed a scalable police wireless system using Node.js, Vue.js, handling 10K+ concurrent connections with RESTful APIs (Express.js) to enable real-time communication and secure data exchange between police officers, and databases.
- Established encrypted communication channels with Node.js, Docker, and HTTPS, ensuring secure data transmission for 1,000+ police officers statewide, protecting sensitive information.
- Integrated WebSocket-based real-time alerting system, enabling instant incident reporting and coordination between officers, reducing emergency response times by 40%.
- Optimized system performance by implementing Redis caching, reducing API response times by 85%, enhancing data retrieval efficiency for frequently accessed information.

Projects

LLM Benchmarking Suite | *PyTorch, Hugging Face Transformers, P-Tuning, CUDA* — Nov. 2024 – Feb. 2025

- Developed a benchmarking framework to evaluate LLMs, optimizing inference speed and memory efficiency with PyTorch and CUDA, giving insights on model selection and deployment strategies for real-world applications.

Tour Management System | *React, Docker, MongoDB, OAuth2.0, Node.js* — Jan. 2024 – May 2024

- Architected a microservices-based tour management application using React and Docker, achieving a 70% reduction in load times by leveraging CDN integration and caching.

Certifications, Publications and Honors

Video-Based Sign Language Translation System Using Machine Learning | *IEEE Publication* — June 2021
Certified AWS Solutions Architect - Associate | *Certificate Credentials* | ID: 45T60EZBCMFQ17 — May 2021
Invited Speaker: Led a two-day Web Development Bootcamp — Sep. 2019

Technical Skills

Programming Languages: Python, Java, C++, C, C#, SQL, JavaScript, TypeScript, PHP, HTML/CSS
Web and Cloud tools: Node.js, React, Angular, FastAPI, Flask, AWS, Azure, Express.js, Redis, RabbitMQ, Webpack
Machine Learning & AI: TensorRT, PyTorch, Hugging Face Transformers, Amazon SageMaker, LangChain, Ollama, CUDA
Data Platform: MySQL, MS SQL Server, Hadoop, MongoDB, SnowFlake, PostgreSQL
DevOps: Docker, GitHub, Git, CI/CD AWS, Gradle, Kubernetes