

EDUCATION

Purdue University <i>Master of Science in Computer Science (GPA: 4.0/4.0)</i> <ul style="list-style-type: none">◦ Research Area: Mechanistic Interpretation of LLMs Efficient LLM Optimizations◦ Coursework: Distributed Systems, Program Analysis, Computer Networks, Information Security, Graduate Algorithms, Reasoning with LLMs[†], ML Systems[†]	West Lafayette, IN 08/2023 - 05/2025 (Expected)
Indian Institute of Technology (IIT), Delhi <i>Bachelor of Technology in Electrical Engineering (GPA: 3.93/4.0)</i> <i>Minor in Computer Science (GPA: 4.0/4.0): [Dean's List]</i> <ul style="list-style-type: none">◦ Publication: A deep learning framework for the detection of tropical cyclones from satellite images. <i>IEEE GRSS</i>◦ Research Thesis: Demographic Prediction from Satellite Imagery using Deep Learning◦ Coursework: Operating Systems, Natural Language Processing, Deep Learning, Computer Vision, System Design Practices, Computer Architecture, Information Retrieval, Machine Learning, Convex Optimization for ML	New Delhi, India 07/2019 - 05/2023

TECHNICAL SKILLS

Programming	C, C++, Java, Python, Go, Javascript, Scala, Rego
Development	NodeJS, React, MySQL, MongoDB, FastAPI, gRPC, Neo4J, Akka, Kafka
DevOps	Docker, Kubernetes, Git, CI/CD (CircleCI), AWS, GCP, BigQuery, Prometheus, Grafana
ML	TensorFlow, Pytorch, OpenCV, LangChain, AutoML, Sagemaker, LangSmith

EXPERIENCE

Deductive AI MLE Internship <i>Topic: Low Latency Ingestion and Code Reasoning Received Return Offer</i> <ul style="list-style-type: none">◦ Worked in a startup environment (<10 members), taking ownership of key features and managing various ad-hoc tasks.◦ Reduced latency of ingestion pipeline from hours to seconds using Akka (Scala) streams and DBMS optimizations.◦ Owned the E2E pipeline for ingestion, retrieval, reasoning & evaluation of code and its interaction with telemetry data.◦ Implemented an agent-less approach which outperforms SOTA (SWE-bench) and cuts down on LLM costs by <i>10x</i>.	Mountain View, CA 05/2024 - 08/2024
DevRev.ai MLE Internship <i>Topic: Incorporating Retrieval Augmented Generation Received Return Offer</i> <ul style="list-style-type: none">◦ Led the development of a retrieval augmented conversational agent RPC from scratch. Implementation fetches neighbors from a vector DB, data from S3 bucket and memory from Redis; currently the top selling feature of DevRev.◦ Implemented a custom <i>LangChain</i>-like solution for chaining LLM calls and added support for function calling. Created an RPC for converting natural language queries into API calls; integrated into a general purpose search bar.◦ Managed the migration of a microservice from Golang to Python to enhance the development of ML serving pipelines.◦ Established a comprehensive encoder benchmarking pipeline using AWS SageMaker for proprietary datasets.	Bangalore, India 05/2023 - 07/2023
DevRev.ai SDE Internship <i>Topic: Adding support for third party integrations Received Return Offer</i> <ul style="list-style-type: none">◦ Contributed to backend in Golang to support 2-way communication with other SaaS apps like Slack and GitHub.◦ Exposed internal workflows through RPCs to enable users to write automations using OPA policy Rego.◦ Worked on multiple integrations for Slack, leading to DevRev's initial breakthrough with their first customers.	Bangalore, India 05/2022 - 08/2022
Sharechat AI MLE Internship <i>Topic: Rule based modelling of DL models for CTR prediction Received Return Offer</i> <ul style="list-style-type: none">◦ Trained a DNN model feeding in a large number of continuous & categorical features for ads CTR prediction; achieved test AUC score of 0.76 on Sharechat's proprietary dataset (3.5% improvement over existing implementation).◦ Formulated RuleNet for distilling rules based on features with historically consistent correlation into models prediction.◦ Productionized model using GCP for serving predictions; setup Airflow job for regular re-training from BigQuery.	Remote (Part-time) 12/2021 - 05/2022

PROJECT HIGHLIGHTS

- **Efficient LLM Optimizations:** Investigating the practicality of low-precision optimizations like QLoRA by comparing their performance in domain adaptation tasks against traditional LoRa for fine-tuning LLaMA-2-7b.
- **Distributed Systems:** Implemented a linearizable sharded key/value storage system using Paxos for fault tolerance and scalability to support cross-group transactions while ensuring robustness against system failures and network partitions.
- **Program Analysis:** Developed a Valgrind tool for dynamic analysis, detecting data dependencies in C code. Also implemented an LLVM module for static analysis, identifying memory leaks in C programs.
- **Vulnerabilities and Attacks:** Investigated vulnerable C codes susceptible to stack-smashing attacks. Attacked diverse vulnerabilities including buffer overflow and DEP bypass while using GDB for debugging and analyzing memory locations.
- **Network Bandwidth Allocation:** Developed and implemented a linear programming solution for optimal bandwidth allocation in multi-stream video analytics (using YOLOv8) within edge computing.
- **Natural Language Inference:** Implemented Few-Shot Cross-lingual transfer learning approaches using Adapter modules and fine-tuned XLM-R models for transferring knowledge from high-resource languages to low-resource languages.