Satya Sai Prudhvi krishna Nikku

☎ 413-406-0703 ✉ snikku@umass.edu linkedin.com/in/prudhvinikku github.com/Prudhvinik1

Professional Experience

Meta **February 2025 – May 2025**

Extern — Capstone Project *Amherst, MA*

- Generated **100K+ synthetic personas** and persona-aware **Math QA pairs** using **zero-shot**, **few-shot**, and **POT prompting** boosting diversity by **10%** via increased unique 1-grams and fine-tuned Llama models using LoRA using HF,wandb.
- Accelerated **LLM inference** throughput by **40%** by engineering an **asynchronous multiprocessing pipeline** integrating **Together AI** and **OpenRouter APIs**, reducing latency for large-scale workloads.

RedBus **June 2022 – July 2023**

Software Engineer – Backend Distributed Systems *Bengaluru, India*

- Designed RedPass **Go** microservice and integrated with **Java/.NET** booking flows; unified opt-in and payment logic, enabling **10K+** daily purchases and **$500K+** annual revenue.
- Maintained and revamped 20+ high-throughput microservices across **Personalization, Search and Booking services written in Java/Go/.NET** handling **3M+ queries per second** for **12M daily users**.
- Delivered a **geospatial "Nearby Boarding Point"** feature in search service used by **50K+** monthly riders.
- Optimized and migrated **RabbitMQ-based schedulers** from Windows to Linux machines, reducing AWS costs by **$600/year** and improving operational efficiency across notification systems.

Tata Consultancy Services **August 2020 – June 2022**

Software Engineer *Hyderabad, India*

- Engineered custom wrapper APIs in **C#** and **WCF**, eliminating redundant XML fields and metadata, reducing average payload size from **40KB to 20KB** (**50% decrease**) and boosting backend data throughput by **10%**.
- Implemented unit and functional **tests** for **15+ features**, boosting release quality by **15%** with QA and product teams.
- Developed **30+ UI pages** in **.NET**, improving operator experience and shopfloor workflow efficiency.

PiChain Labs **January 2020 – June 2020**

Software Engineering Intern *Bengaluru, India*

- Led backend development of a KYC/AML engine for **3+ clients**, delivering **10+ production-grade REST APIs** using **Flask, MongoDB**, and deploying services on **AWS EC2** to enhance compliance and risk mitigation.
- Architected a scalable **Knowledge Graph system** using **Neo4j, Python (Py2Neo, CQL)**, enabling advanced relationship modeling for regulatory entities and improving AML insights.

IBM Research **June 2019 – November 2019**

Machine Learning Intern *Sricity, India*

- Improved aggressive behavior detection accuracy by **5–13%** using deep neural networks and **Convolutional-LSTMs** for spatio-temporal video modeling.
- Experimented **Faster R-CNN** for handgun detection in surveillance videos, to study threat recognition in real-time video analysis.

Projects

Deep Research Assistant | *Next.js, FastAPI, TypeScript, Python* **September 2025**

- Built AI research assistant with FastAPI and Next.js 15, integrating Exa API for web search and Cerebras Cloud (Llama 4) for real-time streaming analysis via Server-Sent Events with sub-second latency
- Developed full-stack TypeScript/Python application with async streaming architecture, implementing concurrent AI inference and markdown rendering across 5+ sources per query

Emotion Cause Pair Extraction (Semeval 2024) | *PyTorch, Python, HuggingFace, Peft* | 🔗 **May 2024**

- Explored and evaluated a question-answering paradigm for ECPE, introducing innovative methodologies that increased emotion-cause pair extraction accuracy by 22% and improved model interpretability.
- Integrated Quantized Low-Rank Adaptation (QLoRA) for efficient fine-tuning of large pre-trained language models, boosting performance by 18% while reducing computational resource usage by 30%.

Deep RL Algorithms Implementation and Evaluation | *Python, openAI Gym, Pytorch* **December 2024**

- Implemented and benchmarked advanced reinforcement learning algorithms (REINFORCE with Baseline, One-Step Actor-Critic, PPO, and N-step SARSA) using PyTorch and OpenAI gym for Policy Optimization

Technical Skills

Languages: Python, Go, Java, C#, JavaScript, Typescript C, SQL, CQL, JSON, YAML

Machine Learning/AI: PyTorch,Scikit-learn, Huggingface, Transformers, vLLM, Numpy, Pandas, LangChain

Web & API Development: Django, Flask, .NET Core, Java REST, HTML, REST API, gRPC, GraphQL, React, Next.js

DevOps & Tools: Docker, Kubernetes, Git, Jenkins, ELK, Postman, AWS, Bitbucket, Jira, GitHub, Unix, Windows, Agile, Cursor

Databases & Messaging: MySQL, PostgreSQL, MongoDB, Redis, Neo4J, Kafka, RabbitMQ, Pinecone, Supabase

Education

University of Massachusetts - Amherst **September 2023 – May 2025**

Masters of Sciences: Computer Science | *CGPA* : 3.82/4.0 *Amherst, MA*

- Coursework: Distributed Systems,Advanced NLP, Systems for Deep Learning, Reinforcement Learning,Software Engineering

Indian Institute of Information Technology, Sricity **July 2016 – June 2020**

Bachelors of Technology: Computer Science and Engineering *Sricity, India*