Yatharth Kapadia

+1 (930) 333-4182 | yatharth.k2@outlook.com | LinkedIn | GitHub | Personal Website

EDUCATION

Master's in Computer Science, Indiana University Bloomington — **Bloomington, USA**
Relevant Courses: Applied Algorithm, Database Design, Elements of AI, Data Mining, Computer Vision, Applied ML, ML in Signal Processing. (GPA : 3.66/4)

Bachelors' in Electronics & Communication, MIT WPU — **Pune, India**
Relevant Courses: Object Oriented Coding, Data structure & Algorithm, Optimization Techniques, Fuzzy logic and Graph Theory, Data Science. (GPA : 3.93/4)

EXPERIENCE

ML Intern, *Outspeed* ***[****Website****]*** 06/2024 – 09/2024 | **San Francisco, USA**

- Led end-to-end fine-tuning and deployment of custom Llama 3.1 LLM using Vllm Inference engine and TensorRT, replacing ChatGPT API dependency and achieving 30% cost reduction while maintaining comparable performance.
- Developed and deployed real-time micro-service for speech-to-text (Whisper) and text-to-speech (Parler TTS) on an AWS EC2 GPU instance.
- Engineered threads queues for parallel communication between STT, TTS, and LLM, streamlining architecture for optimized performance.
- Consolidated all modules (STT, TTS, LLM) on a single GPU, leading to **under 800ms latency** for operations.

Research Assistant, *DSAIL Labs* ***[****Website****]*** 03/2024 –12/2024 | **Bloomington, USA**

- Assessed LLM security using Nvidia Garak, MSCounterfeit, and PurpleLLama, reporting over 15 vulnerabilities.
- Mitigated 5 key AI security risks, **cutting potential exploit vectors by 60%** through targeted guard railing strategies.

Software Developer Intern, *IDeaS* ***[****Website****]*** 07/2022 – 01/2023 | **Pune, India**

- Reduced engineering workload by 100 hours/month by developing a React, Next.js, Express.js, and Plotly-based analytics platform that automatically generates 10 advanced visualizations upon data receipt for global hotel chains, replacing outdated MS Excel macro solutions.
- Collaborated with global stakeholders and product managers to develop an Airflow and Pandas-based data validation pipeline, implementing 8 critical tests to ensure 100% data quality for training, with automated data health reports emailed to the designated officer.

ML Intern, Quidich Innovation Labs ***[****Website****]*** 05/2022 – 06/2022 | **Mumbai, India**

- Designed a predictive algorithm by combining rule-based criteria with player movement trajectories and coordinates to anticipate the moment a bowler releases the ball, **achieving a 95% accuracy rate** and enhancing decision-making in cricket analytics.
- Redesigned the player tracking system in YoloV4 by integrating unique ID assignment and memory retention for each detected cricket player using ByteTrack, significantly enhancing the situational awareness across **3 integrated software systems.**

Software Developer Intern, *Azodha* ***[****Website****]*** 09/2021 – 04/2022 | **New York, USA**

- Led the Open-Notif project, developing a CLI for user-company connectivity with AWS (Lambda, SNS, SQS), MongoDB and Twilio. Implemented Infrastructure as Code by using AWS Chalice to dispatch event-triggered messages, achieving **sub-3-second notification delivery.**
- Architected the Ringisho app's NLP framework, deploying BERT models for phrase suggestion and profanity detection, yielding **92% positive feedback** and **98% accuracy**. Deployed the solution on GCP Compute Engine with Docker, optimizing Golang Socket for real-time processing.
- Crafted a web-based KYC platform leveraging WebAssembly, TensorFlowJS, OpenCV, and JavaScript, **achieving 90% accuracy in client-side identity verification**. Enhanced security and operational efficiency by enabling local data processing, reducing server load and response times.

SKILLS

Language — Python | GO | JavaScript | Java | SQL | TypeScript | GraphQL
FrameWork/Tools — Pytorch | Tensorflow | Django | React (Next.js, Tailwind CSS, NextJS) | Fastapi | Docker | AWS | Kafka | Model-Context-Protocol
DataBase — MySQL | MongoDB | Neo4j | AWS (DynamoDB & RDS) | Redis Caching
Core Competency — ML Model Finetuning | AI Integration in Web Apps & Platforms | AWS Cost Optimization | Full Stack Development
Professional Skills — Stakeholder Management | Client Presentation | Cross-functional Collaboration | Agile Methodologies |

HONORS AND COMMUNITY ENGAGEMENT

Graduate Teaching Assistant (Elements of AI, Prof. David Leake) (ML in Signal Processing, Prof. Jonathan Pooniah) 08/2024
1st prize at Smart India Hackathon by **Dell** 08/2022
4th prize at India Academia Connect AI Hackathon by **Nvidia** 10/2021
87th Rank in Amazon Ml challenge by **Amazon** 08/2021

TECHNICAL PROJECTS

INpersona LLM | Skills: NLP, LLM, FullStack Development [*Demo*] [*Code*]

- Developed an AI assistant using LLaMA 3.2 11B, LlamaIndex, Langchain and ChromaDB with a React (TypeScript, Next.js, Tailwind CSS) frontend, showcasing my professional journey to users through interactive, context-aware conversations.
- **Reduced LLM latency by 40%** with real-time streaming, multi-threaded backend, and dataclass-powered memory for instant, seamless interactions.
- Implemented Redis caching for frequently asked questions, dramatically reducing response latency from **2000ms to 50ms (97.5% reduction)** while **decreasing LLM inference calls by 40%,** resulting in significant cost savings and near-instantaneous user experience.
- Optimized hybrid search system using Knowledge Graph and HyDE, delivering **10% more precise responses.**

FaceInpainting | Skills: Computer Vision, Image Restoration and Algorithm Innovation [*Github*]

- Implemented **Partial Convolution architecture** with PyTorch for image restoration, incorporating a novel approach by leveraging **random walk algorithm** to generate 50,000 specialized binary masks, preventing overfitting.
- Attained an impressive **97% accuracy** in precisely reconstructing severely distorted facial features, with training on a dataset of **50,000** images from **CelebA** highlighting its potential for forensic applications.