




# Yatharth Kapadia

(930) 333-4182   yatharth.k2@outlook.com    LinkedIn    Twitter    Github    Personal Website

## EDUCATION

**Master's in Computer Science** 08/2023 – 05/2025 | Bloomington, USA  
Indiana University Bloomington (GPA : 3.62/4) [[Transcript](#)]

**Bachelors in Electronics and Communication** 08/2019 – 05/2023 | Pune, India  
MIT WPU (GPA : 9.32/10) [[Transcript](#)]

## TECHNICAL SKILLS

**Language** — Python | GOLang | HTML | CSS | JavaScript | Java | SLURM Scripting | SQL

**Technology** — GIT | PyTorch | Tensorflow | Django | NodeJS | ReactJS | TailWind CSS | Fastapi | Docker | AWS | GraphQL | Databricks | Spark | Airflow | GitHub Actions

**DataBase** — MySQL | MongoDB | Neo4j | CassandraDB | DynamoDB

**Core Competency** — ML Model Finetuning | AI Integration in Web Apps & Building AI-Powered Platforms | AWS Cost Optimization | CI/CD

## EXPERIENCE

**Machine Learning Intern** 06/2024 – 09/2024 | San Francisco, USA  
*Ouspeed* [[Website](#)]

- Led the fine-tuning and deployment of an in-house LLM using **Llama 3.1**, reducing reliance on the **ChatGPT API** lowering API costs by 30%.
- Developed and deployed real-time APIs for speech-to-text (**Whisper**) and text-to-speech (**Parler TTS**) on an **AWS EC2** instance with an **A10g GPU**.
- Engineered **threads** 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** 03/2024 – 12/2024 | Bloomington, USA  
*DSAIL Labs, Kelley School of Business* [[Website](#)]

- Conducted security assessments on AI models like Llama and ChatGPT using **OWASP Top 10 criteria**, leveraging tools such as **CounterFIT**, **ProtectAI**, and **Garak** to detect and address vulnerabilities including prompt injection, information disclosure, and hallucinations.

**Associate Software Developer Intern** 07/2022 – 01/2023 | Pune, India  
*IdeaS* [[Website](#)]

- Engineered a **Python and React based workflow platform** using **Seaborn** that auto-generates 10 complex visualizations upon Excel file receipt for global hotel chains, reducing computation times by 20% and saving 100 engineering hours monthly.
- Collaborated with global cross-functional stakeholders and product managers to develop a Python and **Salesforce-based data validation system**, implementing 8 critical tests that ensured **100% data quality for training purposes**.

**Machine Learning Intern** 05/2022 – 06/2022 | Mumbai, India  
*Quidich Innovation labs* [[Website](#)]

- **Led the development** of 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** 09/2021 – 04/2022 | New York, USA  
*Azodha* [[Website](#)]

- **Led the Open-Notif** project, developing a CLI for user-company connectivity with **AWS Lambda**, **AWS Pub/Sub**, **AWS Chalice**, **MongoDB**, and **Twilio**. Implemented 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 **FastAPI** 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.

## HONORS AND COMMUNITY ENGAGEMENT

Graduate Teaching Assistant (Elements of AI under [Prof. David Leake](#)) (ML in Signal Processing under [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, Information Retrieval [[Demo](#)] [[Code](#)]

- Built an AI assistant using **LLaMA 3.2 11B** model and **LlamaIndex**, enabling personalized interaction through fine-tuning, allowing potential collaborators and employers to evaluate professional capabilities through natural conversation
- Established hybrid search combining **Knowledge Graph** and **ChromaDB** with **HyDE** and **Agentic-supervision** architecture, delivering precise, contextual responses while maintaining high accuracy through automated validation

**Facelnpainting** | 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.