Niyathi Allu

California | linkedin.com/in/niyathi-allu | +1 (858) 305-7496 | nallu002@ucr.edu

EDUCATION

MS Computer Science, University of California, Riverside **Sep 2022 - May 2024 (**Expected**) |** USA
Coursework - Advanced Machine Learning, Natural Language Processing, Deep Learning, Advanced Linear Algebra
BTech in Electrical Engineering, National Institute of Technology, Warangal **Aug 2014 - May 2018 |** India
Coursework - Database Management System, Operating Systems, Computer Networks, Data structures and Algorithms

TECHNICAL SKILLS :

- **Programming Languages :** Golang, Python, Java, C++, Nodejs, HTML, CSS, JavaScript, Typescript.
- **Tools & Technologies :** Kubernetes (K8s), Docker, GCP, AWS, Redis, Protobuf, PubSub, Git, Digital Ocean.
- **Machine Learning Tools :** PyTorch, SciPy, HuggingFace, Matplotlib, Scikit-learn, Pandas, Tensorflow, Wandb, OpenAI.

WORK EXPERIENCE

InTheLoop.ai, Machine Learning Engineer Intern **Oct 2023 - Present | San Francisco, CA**

- Engineered and deployed a cutting-edge apparel recognition filter using the **CLIP ViT-bigG-14** architecture, achieving an outstanding **96.6%** accuracy in zero-shot learning.
- This innovation resulted in a marked improvement in user experience, coupled with a substantial reduction in processing time.
- Devised a novel solution to increase the effectiveness in identifying the sub-attributes of the apparel from mere **21.45%** to **76.83%** with fine-tuning of **ViT-bigG-14**, leveraging the image-text model and few-shot learning capabilities with cached key-value pairs.
- Employed AWS S3 as a robust data collection repository and seamlessly integrated AWS SageMaker for streamlined data labeling for the fine-tuning process.
- **Technologies Used :** PyTorch, Tensorflow, OpenClip, MultiModality, AWS S3, AWS SageMaker, VisionModels

University of California, Riverside, Graduate Student Researcher, NLP Labs **Jun 2023 - Present | Riverside, CA**

- Innovated NLP applications, integrating **Vicuna-7B and Falcon-40B** into real-world scenarios, particularly addressing sensitive topics like sexual harassment under Dr. Yue Dong's guidance.
- Orchestrated a robust data pipeline by integrating Vicuna-7B and the FastChat API, facilitating the creation of positive and negative examples tailored for identifying attributes in sexual harassment conversations.
- Applied **contrastive learning techniques** to enhance a chatbot's capabilities, enabling it to discern positive and negative scenarios in sexual harassment conversations and shift towards a more empathetic response.
- Guided project development closely with Dr. Yue Dong, aligning research insights with practical applications for enriched foundations in **empathetic AI solutions**.
- **Technologies Used:** Python, Tensorflow, PyTorch, Numpy, FastChat API, Large Language Models.

Kayapay.ai, Generative AI Engineering Intern **Jul 2023 - Sep 2023 | San Francisco, CA**

- Took the lead in crafting a sophisticated QA chatbot, harnessing the capabilities of OpenAI's GPT-3 model.
- Orchestrated the chatbot's ability to initiate conversations, produce contextually accurate responses, and seamlessly preserve conversation history.
- Led design and development of unified ordering and tracking platform, utilizing Python, FastAPI and Postgresql. Seamlessly integrated Twilio, Slack, & Notion data for optimized workflows. In charge of overseeing deployments on the Digital Ocean.
- **Technologies Used :** Python, Postgres, FastAPI, Celery, Alembic, Docker, GPT 3.5, Pytorch, Digital Ocean, Supabase, Git.

Conversenow.ai, **Software Development Engineer** **Jun 2021 - Aug 2022 | Bangalore, India**

- Designed and Implemented an interface to seamlessly integrate an AI order taking system with Point of Sale terminals, leveraging Golang, Cockroach Database, and GCP, leading to a successfully filed patent.
- Improved synchronization of updates from POS terminals through PubSub and database caching using Redis reducing the API latency from 500ms to 20 ms.
- Minimized operational downtime in active stores without requiring server restarts, utilizing Webhooks, resulting in onboarding 142 stores of Domino's Franchise that accounted for 63% of the successful phone orders.

- Collaborated with the cross functional and customer success teams, and launched features like version control and rollback capabilities within the application, streamlining onboarding procedures and ensuring expedited, efficient deployments.
- **Technologies Used:** Golang, gRPC, Rest API, CockroachDB, Redis, Webooks, GCP, Typescript, Angular, SQL boiler, PubSub.

RetailMeNot Inc, **Software Development Engineer** **Oct 2020 - May 2021 | Bangalore, India**

- Initiated and deployed a sophisticated GraphQL API platform, elevating data retrieval efficiency and bolstering Chrome extension (that supports over 800 websites) capabilities.
- Integrated the Zipkin tracer to seamlessly trace API calls, augmenting visibility and enabling robust monitoring capabilities.
- **Technologies Used:** Python, GraphQL, Grafana, API Gateway, Javascript, Angular, Algolia.

Delta Electronics, R & D, **Assistant Engineer** **Jul 2018 - Apr 2019 | Bangalore, India**

- Engineered a second-order generalized integrator with a phase-locked loop for a 3x130KVA auxiliary converter prototype. Spearheaded software development for remote system control and logging using Hermes.
- **Technologies Used:** C++, Matlab, Postgres.

RESEARCH & PROJECTS

Calc GPT [Code] **Mar 2023 - Jun 2023**

- This project is aimed to understand how the language models perform on mathematical equations and to understand more about domain-generalization issues.
- Conducted an evaluation of the performance of the **EleutherAI/gpt-neo-1.3B** model in calculator development without fine-tuning, revealing inconsistent results on the simple mathematical equations.
- Successfully fine-tuned the **gpt2-medium** model, resulting in a remarkable accuracy improvement from **3.5%** to **79%.**
- **Technologies Used** : GPT2-Medium, PyTorch, Numpy, HuggingFace, Python, Natural Language Processing.

Abstractive Text Summarisation of Legal Documents [Code] **Jan 2023 - Mar 2023**

- Successfully addressed summarization of lengthy documents (5000-6000 words) using the BART model. And Conducted comparative analysis with the leading-edge Pegasus model.
- Innovatively tackled quadratic time complexity in self-attention mechanism by introducing LSG (Local, Sparse, and Global Attention) to the pre-trained BART model. Enhanced efficiency by achieving linear O(n) complexity with LSG Attention.
- **Technologies Used** : BART, Pegasus, PyTorch, Numpy, HuggingFace, Python, Natural Language Processing.

RAG based Question Answering (QA) Bot [Code] **Oct 2022 - Dec 2022**

- Devised a proficient retrieval system employing the HNSWFlat algorithm to extract pertinent passages from an index ensuring that the answers by the bot are in context.
- Developed a QA bot delivering precise PDF-based answers, performance on par with T5 fine tuned model.
- **Technologies Used** : T5, Pegasus, PyTorch, Numpy, Python, HuggingFace, Natural Language Processing.

HACKATHONS :

- Received honorable mention by Microsoft and Llama at UC Berkeley's AI Hackathon for Smart Organiser, a revolutionary project leveraging Large Language Models (LLMs) for multimodal search and intelligent automation**.** Have also developed a plugin for the Notion. **[Code]**
- Secured top honors in Smart India Hackathon, NITW 2017, showcasing our innovative approach and creativity.

LEADERSHIP AND HONORS :

- Awarded **Employee of the Quarter** (January 2022 - March 2022) at Conversenow.ai for seamlessly taking over intern mentoring and driving the development of an AI interface. This initiative resulted in the swift onboarding of the entire Domino's franchise within a month.
- Acknowledged for valuable contributions to the development of Proof of Concepts for ZipKin and API Gateways at RetailMeNot, exemplifying excellence in technological innovation.