Niyathi Allu

California | linkedin.com/in/niyathi-allu | +1 (858) 305-7496 | alluniyathi1997@gmail.com

SUMMARY

Research Engineer working on LLMs, agentic systems, and LLM-as-a-Judge evaluation. Deeply interested in alignment and evaluation research, with hands-on experience building scalable experimentation frameworks and synthetic data pipelines that drive measurable performance gains in real-world AI applications.

WORK EXPERIENCE

ByteDance, Research Engineer - Large Language Models **Sep 2024 - Present | San Jose, CA**

- **Alignment & Prompt Optimization :**
  - Developed domain-agnostic prompt optimization algorithms for applications with sparse, noisy, or inconsistent data, leveraging adaptive context selection, domain-aware memory storage, and genetic routing strategies for the better task alignment and to ensure outputs match user intent.
  - Through iterative experimentation and optimization across multiple heterogeneous domain datasets, achieved an **average 29.4% improvement in accuracy**, while reducing ambiguity and improving response relevance and robustness.
  - Built an **agentic pipeline** along with prompt optimization workflows that enables autonomous reasoning, sequential decision-making, and iterative refinement to improve performance across multi-domain tasks.
  - Synthesized a **multi-turn dialogue dataset in collaboration with the foundational model team**, using dialog inpainting for training large language models, with the goal of improving educational applications across different grade levels.
- **Evaluation Frameworks & LLM-as-a-Judge :**
  - Developed a closed-loop evaluation pipeline where LLMs act as judges to generate and score domain-specific metrics, leveraging chain-of-thought reasoning, self-critique and human feedback to improve task alignment and evaluation robustness.
  - Developing model distillation and supervised fine-tuning pipelines for **LLM-as-a-Judge** using **5k curated synthetic and online data points**, improving evaluation accuracy and efficiency while lowering inference and production costs.
- **Experimentation & Hyperparameter Optimization:**
  - Developed a **Ray Tune–based hyperparameter tuning framework**, including end-to-end online data collection and cleaning, running **100+** parallel HPO experiments, significantly accelerating prompt optimization algorithms by 36.7%.
  - Collected, curated, and maintained **50 benchmarking datasets** and developed an **evaluation framework** to assess and compare prompt optimization algorithms, improving reliability and reproducibility.
- **Technologies Used :** Python, Pytorch, RayTune, Large Language Models, NLP

Resultid.ai, Machine Learning Engineer Intern **Jul 2024 - Sep 2024 | San Francisco, CA**

- Leveraged BERTopic to create efficient **clustering algorithms** and utilized unsupervised learning methods to organize and categorize large sets of text data.
- Successfully developed a pipeline to cluster thousands of reviews into distinct topics, enhancing the ability to identify and address customer concerns, and improving the accuracy and relevance of review categorization, leading to more actionable business insights.
- **Technologies Used :** Python, BERTopic, SpaCy, NLTK, AWS, FastAPI, Large Language Models

InTheLoop.ai, Machine Learning Engineer Intern **Oct 2023 - Mar 2024 | 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.
- 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.
- This resulted in a marked improvement in user experience, coupled with a substantial reduction in processing time from 0.7 sec to 0.25 sec, with a 34% decrease in the response time.

- 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, Clip, S3, SageMaker, Vision Language Models.

University of California, Riverside, Graduate Student Researcher, NLP Lab Jun 2023 - June 2024 | 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 dataset tailored for identifying attributes in sexual harassment conversations.
- Adeptly fine-tuned the Llama-7B model using Hugging Face's peft library, incorporating Quantized Low Rank Adaptation **(QLoRA)**, to deliver empathetic responses to users interacting with the chatbot.
- 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.
- 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.
- **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

EDUCATION

MS Computer Science, University of California, Riverside **Sep 2022 - Jun 2024** | 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

PUBLICATIONS (Under Review) :

- *"Can Linguistically Related Languages Guide LLM Translation in Low-Resource Settings?" (* Submitted to **EACL 2025**.)