GAURAV TADKAPALLY

San Francisco, CA | (213) 913-7899 | gaurav.tadkapally@usc.edu | linkedin.com/in/gauravreddy08 | gtadkapally.com

EDUCATION

University of Southern California California, United States
Master of Science in Computer Science: 3.7/4.0 June 2023 - December 2024

- Served as a Teaching Assistant (TA) for the graduate course Applied Machine Learning for Natural Language Processing (ITP 459)

Vellore Institute of Technology Andhra Pradesh, India
Bachelor of Technology in Computer Science and Engineering: 8.94/10 May 2019 - May 2023

EXPERIENCE

Pitney Bowes Connecticut, United States
Data Science Intern June 2024-August 2024

- Designed agentic code assistant for software testing, leveraging **Speculative Decoding** to accelerate inference speed by 3x and **Abstract Syntax Tree (AST) based retrieval** (tree-sitters) for document indexing **(Cursor Clone: Demo)**
- Leveraged **Direct Preference Optimization (DPO)** and **4-bit QLoRA** quantization to finetune codellama, improving model's generative accuracy by 15% (tested via Mutational Testing)
- Implemented retrieval methodologies (BM25-FTS, Contextual Embedding, and Reranking algorithms) to enhance efficiency and accuracy in retrieving relevant codebase context
- Integrated JaCoCo and Mutational Testing (PIT) to automatically evaluate code coverage & test effectiveness of generated unit tests

MUKHAM Andhra Pradesh, India
Machine Learning Engineer Intern October 2022-May 2023

- Optimized facial recognition model for edge deployment (mobile application), leveraging **knowledge distillation, Post-training Quantization (8-bit quantization) and Automatic Mixed Precision**, decreasing model size by 75%
- Designed a Presentation Attack Detection system (facial spoof detection) utilizing the Lucas Kanade algorithm for motion analysis, achieving a 80% success rate in identifying spoofed faces

MUKHAM Pvt Ltd Andhra Pradesh, India
Research Assistant October 2022 - May 2023

- Developed a UAV-based wildfire detection algorithm utilizing the EfficientNetB0 architecture, incorporating **Neural Architecture Search (NAS)** for model optimization, resulting in a 98% precision rate
- Engineered smart glasses with an Object Detection model (Incremental Learning) for visually impaired, leading 78% accuracy

SKILLS AND CERTIFICATIONS

Languages: Python, TypeScript, JavaScript
ML Stack: PyTorch, Tensorflow, HuggingFace, LangChain, Keras, OpenCV, Scikit-learn, Pandas, NumPy
Tools & Technologies: AWS (Cloud Practitioner), Azure (AI Fundamentals), MySQL, MongoDB, Selenium, Redis

ACADEMIC PROJECTS

Poogle: Perplexity Clone (Demo)

- Engineered a multi-agent web search system with 3 specialized agents, coordinated via shared context memory to decompose tasks, parallelize search, and synthesize high-precision answers
- Improved token efficiency by 65% via ID-based memory referencing and vector-embedded semantic retrieval, enabling scalable, context-aware web search

I made AI play Mafia: A multi-agent asynchronous communication (Demo)

- Developed **asynchronous multi-agent AI system**, enabling structured communication among 6+ autonomous agents in social deduction gameplay scenarios
- Implemented modular two-part brain architecture (Scheduler & Generator), with a **concurrency-safe** shared context

AK15: Agentic Kubernetes Middleware (Github)

- Devised an LLM-based middleware that automates Kubernetes cluster read queries, achieving a 93% reduction in contextual token usage through agentic function calling and context retrieval
- Implemented 15 specialized API functions enabling the LLM to perform human-like, context-aware interactions with Kubernetes, optimizing and reducing API costs by leveraging targeted data retrieval strategies

PUBLICATIONS

- Sethuraman, S. C., Reddy Tadkapally, G. et al. **Simplymime: A dynamic gesture recognition and authentication system for smart remote control**. IEEE Sensors Journal (2024). https://doi.org/10.1109/JSEN.2024.3487070
- Sethuraman, Sibi C., Gaurav Reddy Tadkapally, et al. ***iDrone: IoT-Enabled Unmanned Aerial Vehicles for Detecting Wildfires Using Convolutional Neural Networks***. Springer Nature Computer Science (2022). https://doi.org/10.1007/s42979-022-01160-7