Teja Gollapudi

teja@tejagollapudi.com ✻ 3126841770 ✻ Palo Alto, CA
GitHub ✻ Medium ✻ Website ✻ LinkedIn

Work experience

Meta ***Jan 2024 - current***
Senior Research Engineer *Burlingame, CA*

Post Training Foundational Models

- Currently working on inducing and enhancing agentic capabilities in Foundational Models via Reinforcement Learning (GRPO, Reinforce, ARPG).
- Advanced agentic capabilities, reasoning, and factuality in large language models, including tool-calling across multi-turn, multi-tool, and multi-modal scenarios
- End-to-end ownership of LLM design, training, evaluation, and deployment for wearables AI models
- Published research on improving LLM reasoning and factuality at internal and external venues

Wearables Product Launches

- Led AI development for sports features on wearables, driving real-time integrations for major campaigns (Super Bowl 25, Olympics, FIFA CWC)
- Enabled new AI-native product launches including RBM displays and Oakley sports glasses
- Led the development of fitness AI features for workout history on wearables

VMware Inc. ***2021 - 2024***
Machine Learning Engineer *Palo Alto, CA*

Large Language Models (LLMs)

- Co-led development of commercially viable instruction-following open-source LLMs, publishing models on Hugging Face and GitHub
- Fine-tuned code LLM for an internal code completion co-pilot service
- Designed experiments, data-curation strategies, and fine-tuning pipelines for both low-resource hardware and multi-node cloud GPU training
- Provided technical guidance to teams across the organization on adopting LLMs

LLM API

- Built internal LLM API service and Python client as a self-contained AI service, handling **2,000+** daily requests across **40+** teams
- Scaled production systems with vLLM and Ray for enhanced throughput

NLP Models and Optimization

- Fine-tuned RoBERTa and distilled into TinyRoBERTa, achieving **600%** throughput increase and **50%** latency reduction
- Trained multilingual ColBERT model achieving **7x** higher Recall@5 over previous search solution
- Pre-trained domain-specific BERT and distilled into MiniLMv2, open-sourcing both libraries

Leadership and Community

- Founded VMware's Hugging Face page, driving monthly downloads into the **five-digit** range
- Published technical blogs and papers at internal and external venues
- Mentored engineers transitioning into ML roles; conducted interviews for ML positions

Machine Learning Engineering Intern ***Jun 2020 - Jan 2021***
VMware, Inc. *Palo Alto, CA*

- Created an ML spell checker using Seq2Seq transformer architecture

Machine Learning Intern ***May 2018 - Jul 2018***
Inkers.ai *Bengaluru, India*

- Designed a tree-structured Convolutional Neural Network for object detection using Multi-task learning
- Constructed a large image dataset by scraping multiple image sources and using the data to train multiple CNN architectures

Technical skills

Domains	Machine Learning, Natural Language Processing, Generative Models, Computer Vision, Reinforcement Learning
ML Libraries	PyTorch, TorchScript, Tensorflow, Keras, Hugging Face libraries, Numpy, Deepspeed, ONNX Runtime
Languages	Python (Preferred), Java, C++ (Familiar), Javascript (Familiar)
Other	Docker, Kubernetes, FastAPI, SQL, MongoDB, Cloud Platforms (GCP, AWS S3), Ray

Education

Master of Science in Computer Science *University of Illinois at Chicago (UIC)*
Aug 2019 - May 2021 *4.0 GPA*

B.Tech in Information Technology *Vellore Institute of Technology (VIT), India*
Jul 2015 - May 2019

Publications

- TruthRL: Incentivizing Truthful LLMs via Reinforcement Learning (Accepted at ICML 2026), https://arxiv.org/a
- PrismRAG: Boosting RAG factuality with distractor resilience and strategized reasoning, M Kachuee, **T Gollapudi**, EMNLP 2025, https://aclanthology.org/2025.emnlp-industry.53/
- The Unreasonable Effectiveness of Eccentric Automatic Prompts, Rick Battle,**Teja Gollapudi**, 2024, Arxiv, https://arxiv.org/abs/2402.10949
- Fast Multi-scale Face Detection CNN, **N. S. Gollapudi** doi 10.1109/ViTECoN.2019.8899616

Miscellaneous

- **Work featured in News and Journals**: IEEE Spectrum, The Register, Business Insider and more.
- **Blog** : https://medium.com/@tgollapudi10