Kiran Ramnath

LinkedIn | Google Scholar | Homepage

Education

University of Illinois, Urbana-Champaign — Champaign, IL, USA
Master of Science in Electrical and Computer Engineering, GPA: 4.0/4.0 — *Aug. 2019 – May 2021*

Birla Institute of Technology and Science, Pilani — Pilani, India
Bachelor of Engineering in Electrical and Electronics Engineering, GPA 8.55/10 — *Aug. 2012 – Jun. 2016*

Research Experience

Advisor: Prof. Mark Hasegawa-Johnson, Statistical Speech Technology (SST) Group, UIUC
Research Interests: Natural Language Processing, Computer Vision, Speech Processing, Knowledge Graphs
Thesis: Fact-based Visual Question Answering using Knowledge Graph Embeddings

- **PAFT: A Parallel Training Paradigm for Effective LLM Fine-Tuning** *Shiva Kumar Pentyala, Zhichao Wang, Bin Bi,* ***Kiran Ramnath****, Xiang-Bo Mao, Regunathan Radhakrishnan, Sitaram Asur, Na (Claire)Cheng* submitted to ICLR, 2025
- **A Comprehensive Survey of LLM Alignment Techniques: RLHF, RLAIF, PPO, DPO and More** *Zhichao Wang, Bin Bi, Shiva Kumar Pentyala,* ***Kiran Ramnath****, Sougata Chaudhuri, Shubham Mehrotra, Zixu (James)Zhu, Xiang-Bo Mao, Sitaram Asur, Na (Claire)Cheng*, arxiv.org/abs/2407.16216
- **Worldly Wise! Cross-lingual knowledge fusion for Fact-based Spoken Visual Question Answering** ***Kiran Ramnath****, Leda Sari, Mark Hasegawa-Johnson, and Chang D. Yoo.* NAACL-HLT, 2021
- **Seeing is Knowing! Fact-based Visual Question Answering using Knowledge Graph Embedding** ***Kiran Ramnath*** *and Mark Hasegawa-Johnson.* arxiv.org/abs/2012.15484
- **Performance Improvement of Operational Amplifiers in Subthreshold Region** ***Kiran Ramnath*** [1]*, Deepansh Dubey* [1]*, Anu Gupta.* 12th IEEE INDICON (India Conference), 2015
- **Systems and methods of Retrieval Augmented Generation of texts and actions, USPTO 18749760**
- **System for finetuning tenant-specific LLMs for improved retrieval augmented generation of service replies, USPTO 18927209**
- **System for user-feedback based prompt tuning for Generative AI applications, USPTO 18907005**

Work Experience

Applied Scientist 2 — Aug 2024 – Present
AWS Bedrock — *2795 St Austine Drive, Santa Clara, California*

- Core Science team member that delivered Automatic Prompt Optimization (APO) on AWS Bedrock (announced at AWS Re:Invent 2024). APO optimizes user prompts across Amazon Nova, Claude Sonnet 3.5, Claude Sonnet, Claude Opus, Claude Haiku, Llama 3 70B, Llama 3.1 70B, Mistral Large 2, and Titan Text Premier models. **Press link 1**,**Press link 2**, **Announcement**

Senior Applied Scientist — Dec 2023 – July 2024
Salesforce Einstein — *415 Mission St, San Francisco, California*

- Worked on in-house finetuned LLM as AutoEvalutor platform to benchmark, monitor, and improve Einstein Copilot and RAG performance across the Salesforce Agentforce GenAI platform. Our trained model
- Worked on Salesforce's flagship Generative AI offering Einstein Service Replies used by 100+ large enterprise customers. Built novel querying+indexing methods for multi-tenant multi-lingual RAG systems; achieved 2X improvement for Recall@5 and 30% improvement for Recall@10 over baseline.
- Conducted research in LLM alignment that led to research published by dozens of leading AI research groups around the world, and our LLM topping Huggingface OpenLLM leaderboard in April 2024

Senior Member of Technical Staff — Oct 2022 – Nov 2023
Salesforce — *415 Mission St, San Francisco, California*

[1] denotes equal contribution

- **Generative AI:** Led a cross-functional effort with Salesforce AI research team to deploy a Retrieval-Augmented Generative AI (RAG) model for internal developer support, serving close to 10,000 developers. Obtained weak-supervision based on past support interactions to train a sentence transformer retriever using Contrastive Learning. Improved top3 search accuracy by 18% connecting Confluence, Git, StackO, & Slack as doc-sources

Member of Technical Staff July 2021 – Oct 2022

Salesforce *415 Mission St, San Francisco, California*

- **Slack-first AI for customer service:** As the founding data scientist with Hyperforce Development Platform Support, I led several innovative NLP/DS use-cases from prototype to production. These include customer request categorization using Bertopic, operational oncall dashboards, support traffic dashboards / forecasts, etc. providing critical visibility to SVPs and EVPs.
- Was a core contributor to the conversational support bot used in over 100+ internal Slack channels

Infrastructure Data Scientist Intern May 2020 – Aug. 2020

Salesforce *Urbana, IL*

- **Long-term infrastructure demand forecasting.** Built segmentation + forecasting models to predict Salesforce's global infrastructure requirements in 3-5 years. Automated data pipelines for 30+ models

Graduate Teaching Assistant Aug. 2019 - May 2021

University of Illinois, Urbana-Champaign *Champaign, IL*

- **Artificial Intelligence (CS 440 / ECE 448)**: Was teaching assistant for a senior / graduate level course on AI introducing **350+** students to ML, CV, NLP, etc. Held weekly tutorial sessions and office hours, designed coding assignments, autograders, exam problems

Experienced Associate (Select projects) Jun. 2016 – Jun. 2019

PwC US Advisory, BG House, Lake Boulevard Road, Hiranandani Gardens *Powai - 400076,Mumbai, India*

- **Promotion campaign effectiveness.** Identified effectiveness of promotion campaigns for a leading animal-pharma company, helping it shut down 10M$+ non-performing programs. Leveraged linear regression and Bayesian networks in R to perform uplift analysis.
- **Complaints management using text analytics.** Revamped consumer complaints categorization against financial institutions into actionable groups using topic modelling, full-text search engines, sentiment analysis, etc.
- **Frequent itemset mining.** Using Bayesian networks and graph mining in R, designed a novel frequent itemset mining technique based on customer transaction data for a technology specialty distributor, lifting sales by 4%
- **Collection efforts optimization using decision trees.** Built a decision tree model using SAS to help a mortgage client deprioritize accounts & optimize collection efforts by 10% points
- **Resumé shortlisting automation.** Built a resume shortlisting model in Python using random forest text classifier to enable internal robotic process automation. Reduced man-hours spent by 50%
- **Event detection using Named Entity Recognition.** Identified potential healthcare market monopolization through company mergers. Led a team of interns to use named entity recognition on scraped articles to create a PoC by partnering with SMEs

Invited Talks

- **Visual Question Answering: An Overview** Future of Privacy Forum AI Working Group
  ***Kiran Ramnath**, and Mark Hasegawa-Johnson.* 11 Jan, 2021
- **Data Science embedded Management Consulting** SP Jain Institute of Management and Research, India
  ***Kiran Ramnath**, Vidhi Tembhurnikar, and Pradnesh Deshmukh.* June 2019

Courses (* denotes A+)

UIUC: Pattern Recognition, Artificial Intelligence*, Random Processes*, Distributed Systems*, Computational Inference and Learning, Learning-based Robotics, Statistical Inference for Data Scientists and Engineers
Coursera: Design and Analysis of Algorithms - 1 and 2, Machine Learning, Deep Learning Specialization, Natural Language Processing with Attention Models

Technical Skills

Languages: Python, C, R, MATLAB, Java (familiar)
Deep Learning / AI Frameworks: PyTorch, Tensorflow, Keras, PyRobot,Langchain, OpenAI
Developer Tools: Git, Docker, Kubernetes, Shell, Google Cloud Platform, AWS, Selenium, Jenkins, Spinnaker
Web development: HTML, CSS, JavaScript, Django
Data analysis and databases: Tableau CRM, Tableau, Neo4J, SAS, AnyLogic, SQL, SAQL, MS Excel