Kiran Kannar

kiran.kannar@gmail.com • +1 - (858) 405 6676 • LinkedIn/kannarkk

Skilled ML Engineer adept in software development and applying ML/NLP to tackle complex challenges; Strong ownership mentality with a proven track record of building scalable and efficient ML systems; Psychological safety champion.

Education

M.S. in Computer Science **Sep 2016 – Jun 2018**

University of California San Diego; GPA = 4

- **Thesis**: Exploiting Geographical and Temporal Patterns for Personalized POI Recommendation; Advised by Dr. Julian McAuley

B.E. in Computer Science **Sep 2010 – May 2014**

R.V. College of Engineering, Bangalore; GPA = 9.84/10

- *Second Rank Honours* in the graduating batch of 142 students

Work experience

Infinitus Systems **Aug 2022 – Present**

Tech Lead Machine Learning (current), Senior SWE-ML (1.5 years) *San Francisco*

Model Platform Innovation

- Leading Model-as-a-Service design towards optimized, scalable inference for conversational AI
- *Shadow mode*: Led a cross-functional initiative to enhance model iteration and deployment efficiency with non-disruptive testing of new models in live phone calls.
- *Caching*: Designed and built model inference caching system, which cut latencies by up to 37%, boosting automation and minimizing manual intervention.
- *Training*: Cut model training times by 33% by orchestrating a training workflow using Docker Compose and concurrency.

Modeling

- *Prompt modeling*: Introduced prompt models for bad data identification with Gemini Pro on Vertex AI; established best practices for iterative prompt tuning.
- *BERT embedders*: Conducted comprehensive experiments with BERT embeddings for intent classification and entity recognition (1.5% accuracy gain); helped prioritize GPU-based training workflows.
- *Experiments*: LLM Agent for Inbound call automation with Gemini & Langchain; Leverage phonetic signals (G2P) to handle STT mistranscription errors; Multi-choice QA for output field extraction at the end of phone calls.

Technical Excellence

- *Increased automation*: Conversational AI improvements towards 90% IVR, Hold and Auth call phase automation with no human intervention
- *Consistency*: Standardized new model integration with singleton patterns, tracing, and caching integration.
- *Robustness*: Instituted engineering best practices like functional testing, post-mortems, and maintainability of tests.

Leadership

- *NLP Reading Group*: Facilitated discussions on state-of-the-art models, and promote a commitment to continuous learning and innovation.
- *Cross-functional Collaboration*: Engaged with stakeholders including Product, Customer Enablement, and Computational Linguists, optimizing customer impact and internal efficiency
- *Hiring*: Contributed to team growth by interviewing 50 candidates for key ML, NLP and Data roles.
- *Psychological Safety*: Championed psychological safety, leading assessments and advocacy across teams.

Salesforce **Aug 2018 – Aug 2022**

Senior Member of Technical Staff (2 years), Member of Technical Staff (2 years) *San Francisco*

Einstein Agent

- Spearheaded the no-downtime migration of ML apps (Case Classification, Case Wrap-up) to an advanced ML platform, enhancing performance and scalability
- Directed machine learning application health monitoring initiatives across Service Einstein teams; establishing SLI/O metrics, alerts, and dashboards for preemptive issue resolution.
- Developed a proof of concept for live chat summarization during case wrap-ups; architected the project's pilot phase.

Advanced Preventive Maintenance

- Developed a robust system with an innovative DB schema for work order management of maintenance plans and assets with recurring maintenance schedules.

- Analyzed customer data to pinpoint top-20 candidates for targeted required product recommendations on work orders; Identified features for model development and analysis.

Salesforce **Jun 2017 – Sep 2017**
Software Engineering Intern *San Francisco*
- Engineered milestone trackers within Salesforce Lightning, to enhance project management capabilities.

Oracle **Jun 2014 – Jun 2016**
Software Engineer *Bangalore*
- Led accessibility enhancements for PeopleSoft Internet Architecture, significantly improving diverse user group experience.
- Resolved critical development bottlenecks in accessibility; led training sessions to embed accessibility principles early in the development process

PayPal **Jan 2014 – Jun 2014**
Software Engineering Intern *Bangalore*
- Developed Map-Reduce jobs for efficient processing and storage of user click-stream data across distributed data stores.
- Built a Scalable Query Framework (SQF) with HBase and Elasticsearch, delivering near-real-time responses
- Designed and implemented a real-time analytics and visualization interface using D3.JS, NodeJS, and AngularJS, enabling immediate insights into user behaviors.

Research Experience

Sequential Recommender Systems | *UCSD* **Sep 2017 – Jun 2018**
- *MS Thesis* under Dr. Julian McAuley. Leveraged temporal and geographical patterns in human mobility to personalize location recommendations.

Modeling the Evolution of User Expertise | *UCSD* **Apr 2017 – Jun 2017**
- *Independent Project* with Dr. Julian McAuley. Replicated study on expertise evolution in online reviews using latent factor models, verifying patterns in expert and novice rating behaviors.

Project Highlights

Personalized Next Song Recommendation | *Human Behavioral Modeling, UCSD* **Nov 2017**
- Implemented a personalized next-song recommendation system using metric embeddings over user's listening history extracted from 30 Music and Now Playing datasets.

Duplicate Question Detection | *Neural Networks for Pattern Recognition, UCSD* **Mar 2017**
- Evaluated multiple deep neural network models to identify duplicate questions within the Quora dataset, with bi-directional LSTMs, Siamese networks, and pre-trained word embeddings (Word2Vec, GloVe); 81.46% accuracy and 0.755 F1 score with BiLSTMs and GloVe

Bayesian Personalized Ranking (BPR) | *Web Mining and Recommender Systems, UCSD* **Mar 2017**
- Demonstrated the effectiveness of BPR-MF algorithm over traditional collaborative filtering techniques on Yelp data.

Certifications

Natural Language Processing (xCS224n) | *Stanford* **Dec 2020**
Deep Learning (5 courses) | *DeepLearning.AI (Coursera)* **Jul 2020**

Relevant Skills

Languages: Python, Java, C++, SQL, JavaScript, Go
Machine Learning/Deep Learning: PyTorch, TensorFlow, Huggingface, TorchServe, NVIDIA Triton
Data Engineering: Hadoop, HBase, Elasticsearch, Redis, BigQuery, Google PubSub
Cloud & Deployment: GCP, Docker, Vertex AI, Kubernetes
Tools & Monitoring: Git, Prometheus, Argus, Grafana, OpenTelemetry

Awards and Honors

- 4th Place, Cohere Build Day SF - "Mental Health Assistant LLM Agent" (2024)
- Service Einstein Super Star for Trust and Innovation (2021)
- Top Performer, Advanced Preventive Maintenance, awarded bonus stock in 2020 annual review (2020)
- Masters Award for Excellence in Service & Leadership by the Department of CSE, UCSD (2017)
- Second Rank Honours, B.E. in CSE at RVCE (2014)
- Infineon India Merit Scholarship recipient for first rank honors in sophomore year at RVCE (2012)