Kiran Kannar

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

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;
Research Interests: Interpretability, reasoning, alignment, human behavioral modeling, and NLP for specialized domains

Work experience

Infinitus Systems **Aug 2022 – Present**
Staff Machine Learning Engineer (current), Senior MLE (1.5 years) *San Francisco*
Current project areas: AI Safety [AI review, Adverse Event Detection, Clinical Safety Escalations]

AI Review (technical report)

- Developed an ML auto-review system for healthcare voice agents, scaling to **30+ customers in 6 months** and reducing **human review time by 60%+**.
- Developed a modular, observable platform with unified pipelines and multi-modal evaluation (text+audio) framework, leveraging containerized training workflows (Docker Compose) for enhanced reproducibility and security.
- ***Research***: Multi-task rationale-assisted knowledge distillation for AI Review; Retrieval-augmented SOP Adherence evaluation system

Adverse Event Detection (blog)

- **0-to-1**: Developed SAGE (Safety Adverse-event Guidance Engine) - a multi-modal (text+audio) NLP system to detect adverse events on healthcare voice agents with **98% recall** and **0.3% false positive rate** (in production).
- ***Research***: Concept-activation vectors (CAV) for model interpretability and improved precision (blog)

Dialogue Breakdown

- Productionized a dialogue breakdown system with RoBERTa-based breakdown model, heuristics, and VertexAI containerized deployment to determine the need for human assist.
- Led the data labeling strategy, collaborating with annotators to create fine-grained breakdown labels and improve downstream model performance.
- ***Research***: Quantization (PTQ) for model inference optimization; LLM-based replacement models.

Automation Platform

- Investigated and evaluated scalable MaaS architectures (TorchServe, Nvidia Triton, Vertex AI) for NLP model deployment, establishing core Vertex AI deployment patterns adopted across company
- Led conversational AI initiatives towards **90%** automation of phone calls with standardized model integration (singletons, tracing, caching) and engineering best practices (testing, post-mortems).
- Architected and deployed model inference caching that realized a **33%** latency reduction, alongside a non-disruptive shadow testing framework for live calls to accelerate model iteration velocity
- ***Research***: 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 and Leadership

- Initiated and led *ML Review* sessions - a cross-team forum for rigorous evaluation of modeling, service design and MLOps practices, and cross-pollination of ideas.
- Drove cross-functional collaboration (Product, Customer Enablement, Linguistics) optimizing customer impact and efficiency.
- *NLP Reading Group*: Facilitated discussions on state-of-the-art models, and promote a commitment to continuous learning and innovation.
- Fostered team growth and psychological safety through mentoring junior ML engineers, hiring (50+ interviews) and advocacy efforts.

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
- Developed a proof of concept for live chat summarization during case wrap-ups; architected the project's pilot phase.
- Directed machine learning application health monitoring initiatives across Service Einstein teams; establishing SLI/O metrics, alerts, and dashboards for preemptive issue resolution.

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 orgs 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 in core architecture, directly addressing critical development bottlenecks and fostering inclusive design practices; India and China Accessibility Lead.

PayPal **Jan 2014 – Jun 2014**
Software Engineering Intern *Bangalore*

- Built a Scalable Query Framework (SQF) with HBase and Elasticsearch, delivering near-real-time responses with Map-Reduce jobs for efficient processing and storage of user click-stream data across distributed data stores.
- 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. (thesis)

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.

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
- Masters Award for Excellence *in Service & Leadership*

B.E. in Computer Science **Sep 2010 – May 2014**
R.V. College of Engineering, Bangalore; GPA = 9.84/10

- **Thesis**: Scalable Query Framework for Near-Real Time Responses; Advised by PayPal
- *Second Rank Honors* in the graduating batch of 142 students

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.

Relevant Skills

Languages: Python, Java, C++, SQL, JavaScript
Machine Learning/Deep Learning: PyTorch, TensorFlow, XGBoost, 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