KANCHAN ASHOK NAIK

+1 (408)-690-4186 | San Jose, CA | kanchannaik55@gmail.com |linkedin.com/in/kanchan-ashok-naik| https://github.com/kanchNaik

*AI-focused Software Engineer with **5+ years** experience delivering global-scale products and **AI-powered features**. Proven track record in **microservices, distributed systems, and full-stack delivery**, combined with expertise in **LLM/RAG pipelines, embeddings, and scalable inference APIs**. Skilled at building **production-ready AI integrations** that improve latency, reliability, and user experience at enterprise scale.*

TECHNICAL SKILLS

Languages: Python, C#, JavaScript, Java, SQL
Backend: .NET Core, Django, Flask, FastAPI, REST APIs, Microservices
Frontend: ReactJS, Redux, HTML5, CSS3, Bootstrap
Databases: PostgreSQL, MySQL, MongoDB, Redis
Cloud/DevOps: AWS, GCP, Docker, Kubernetes, Git, CI/CD, MLflow
AI/LLM Tools: OpenAI, Hugging Face, Sentence-BERT, CLIP/BLIP, Pinecone, FAISS, ChromaDB, RAG pipelines, LLMOps, Prompt Engineering

WORK EXPERIENCE

AI/ML Engineering Intern

Pyramyd.ai (May 2025 – Aug 2025), Berkeley, CA

- Built an AI requirement-extraction & scoring engine ranking **~3K** SaaS vendors; deployed as a backend service powering "Top Vendors" widget for GTM teams.
- Shipped a **real-time RAG** search API (Pinecone/FAISS) with **P95 latency less than 300ms** and **99.9% uptime**, enabling enterprise buyers to compare vendors instantly.
  Automated **model refresh** and **fairness checks** via CI/CD pipelines (**MLflow** + GCP), eliminating manual retraining overhead.
- Designed capability-coverage clustering pipeline (**UMAP**, **HDBSCAN**) reducing solution architect analysis time by **3.3x**.

Senior Software Engineer

Musafir.com India Pvt. Ltd. (Jul 2018 – Nov 2023, India)

- Drove architectural migration from monolithic .NET to **.NET Core** and **Python microservices,** cutting **API latency 50%**, downtime 40%, and scaling to 1M+ daily requests.
- Architected core booking microservices (flights, hotels, packages), improving modularity, reducing search-to-booking **latency 45%,** and enabling new revenue streams.
- Engineered **hybrid recommender** and **fraud-detection** pipeline that lifted booking **conversions 25%,** click through rate **30%**, and blocked 80%+ bot traffic, saving $600K annually.
- Optimized performance with **async ops**, Redis caching, & DB query tuning, boosting **throughput 50%** & cutting peak **response time 60%**.
- Integrated 10+ global airline & hotel APIs with SLA management, retries, and error recovery, ensuring **99.95% uptime** for real-time booking.
- Rebuilt finance workflows (invoicing, refunds, reconciliation) in **microservices + PostgreSQL**, reducing reconciliation effort from **3 days** to **1 day**.
- Designed **Kafka** and **Spark ETL** with **Looker dashboards**, shrinking reporting lag from **24h** to **5min** for P&L and booking insights.
- Delivered NLP "text-to-booking" solution during hackathon, enabling SMS-based flight booking and restoring **$100K/month** revenue during supplier API outage.
- Mentored and led 12-member engineering squad, setting technical direction, reviewing designs, and accelerating release cadence by 30%

AI & SYSTEMS PROJECTS

Distributed-Hailing Platform
ReactJS, Django, MongoDB, AWS, Kafka, PySpark, Redis

- Engineered a real-time ride-matching backend handling **10K+ concurrent users**, with modular frontend in React and deployed on AWS.
- Optimized query latency using **geospatial indexing** with **Redis caching**, boosting throughput by **40%**.
- Added ML-based dynamic **fare prediction** (Streamlit + Python) to personalize pricing and improve booking trust.

LLM-Based Interview Simulation with RAG & Multimodal Evaluation
- *GPT-4o, LLaMA-2, Mistral-7B, Qwen, ViT, Whisper, Praat, RAG, GCP, Pinecone*

- Built an AI interview simulator retrieving questions from **157K** Q&A **embeddings** (ChromaDB + Sentence-BERT).
- Deployed fine-tuned LLaMA & Mistral models behind **FastAPI** microservices, enabling adaptive Q&A at scale.
- Integrated **Whisper speech-to-text** and **prosody analysis**, supporting real-time voice interviews with less than **350ms** response latency.

Wikinews Insights: Real-time News Trend Detection & Summarization

- Built a **Kafka** and **PySpark** streaming pipeline to process **10M+ hourly** Wikipedia pageviews and correlate with live news headlines.
- Implemented **DGIM** and **exponentially decaying window** algorithms for trend detection and Longformer-Encoder-Decoder (LED) for abstractive summarization, reducing manual article review effort **by 40%**.
- Designed a news classification pipeline (Naïve Bayes, SVM, BERT), with BERT achieving **89% accuracy** and improving **macro-F1 by 12%** over traditional models.

EDUCATION

- **MS in Applied Data Science San Jose State University, San Jose, California, USA,** Expected Dec 2025, GPA 3.87/4.0
  Relevant Coursework**:** Distributed Systems, Deep Learning, Generative AI, Machine Learning, Big Data Technologies
- **B.E.(Information Technology) VIIT, Pune University, India** 2014 - 2018, GPA 3.71/4.0

CERTIFICATIONS and PUBLICATIONS and AWARDS

- **Certificates** - Databricks – Academy Accreditation: Generative AI Fundamentals *Mar 2025 – Mar 2027 · Cred ID: 138825169*
- ***Publications -*** *Fake Currency Detection Using Image Processing and Random Forest Algorithm*, Recent Trends in AI & Its Applications (e-ISSN: 2583 -4819)