Avni Kothari

AI/ML Engineer - Healthcare

Oakland, CA | avni510@gmail.com | +1-630-701-4490

LinkedIn | Github | Website | Google Scholar

- AI/ML Engineer with 5+ years of experience architecting and deploying end-to-end ML and LLM powered production systems in healthcare.
- Specialist in leveraging LLMs and multimodal patient data, and collaborating with clinicians to lead healthcare AI projects from initial conception and research through to scalable, maintained deployment, delivering measurable improvements in clinical workflows.
- Published researcher at top conferences **(ICLR, NeurIPS, ML for Health)** with corresponding open-source software packages.

SKILLS

- **Languages:** Python, TypeScript, Java, Elixir, Clojure
- **AI/ML:** GenAI (GPT 4o, Claude Sonnet 4, Cohere Command A), Langchain, HuggingFace, PyTorch, LLM Evaluations, AI Research, Benchmarking LLMs, Scikit-Learn
- **MLOps:** ML Pipelines and ETL, CI/CD, AWS (EC2, S3, Terraform), Docker, Elasticsearch
- **Software & Database:** REST APIs, GraphQL, SQL (Postgres, DuckDB), React, Flask, Ruby on Rails

WORK EXPERIENCE

AI/ML Engineer; University of California, SF; CA Sept 2023 – Present

- Architected and deployed clinical AI systems to **save social workers 300+ hours/ week**, **identify patients at risk for readmission**, and create simple, interpretable **models from multimodal patient data**
- Drove end-to-end architecture of a GPT 4o powered **clinical summarization pilot to condense patient charts (20K+ tokens) into actionable summaries**, co-developed with 5+ clinicians to improve patient care
- Established and led human and automated annotation pipelines with active learning to **ensure 90%+ accuracy on 1000+ GPT 4o clinical summaries**
- Designed and deployed a scalable ETL pipeline to process health record data, **enabling ML training and evaluation for 3000+ patient records and 30000+ patient visits**
- Engineered, deployed, and evaluated a **custom readmission risk prediction model,** adopted by 10+ clinics, with **12% higher accuracy than the general model**
- Created and presented a hybrid LLM-empirical framework to **reconcile clinical LLM hallucinations, achieving 40% interpretability improvement** (validated by clinical experts) while matching top predictive performance and enabling production deployment
- Developed and authored a publication for a multimodal **ML pipeline to integrate tabular and unstructured clinical data to improve black-box model interpretability**, delivering 7% AUC improvement in readmission risk prediction
- Developed and authored a publication for a **scalable agentic pipeline** using GPT-4 with statistical learning to generate **interpretable clinical features from multimodal data**, outperforming black-box models while maintaining interpretability for decision support

Machine Learning Engineer; University of California, SD; CA June 2022 – Mar 2024

- Developed a model-agnostic bias auditing framework surfacing bias in up to 30% of lending datasets across 5+ models and 3+ datasets with 25+ real-world constraints

PUBLICATIONS

Bayesian Concept Bottleneck Models with LLM Priors
Jean Feng, Avni Kothari, et al; *NeurIPS under review, 2025*

Prediction Without Preclusion: Recourse Verification With Reachable Sets
Avni Kothari, et al; ***ICLR – Top 5% among submissions**, 2024*

Implementing a Predictive Model to Reduce Hospital Readmissions in a Safety Net Healthcare System
Arturo Gasga, Avni Kothari, et al; *ML4H - Machine Learning for Health, 2024* ***Oral Spotlight***

OPEN SOURCE SOFTWARE

Developed Responsible AI methods and software to improve model transparency and safety, architecting both a multimodal, LLM-Bayesian framework for discovering interpretable clinical rules and a model-agnostic audit to identify fairness and preclusion risks.

EDUCATION

University of California, San Diego
Masters in Computer Science
DeepMind Fellow
Thesis: Foundations of Model Agnostic Recourse Verification
San Diego, CA — 2021 - 2023

University of Texas at Austin
Bachelors in Mathematics and Economics
Minor: Computer Science
Austin, Texas — 2011 - 2016

- Collaborated with a lending expert to define real-world constraints on lending datasets to identify individuals subject to preclusion
- Authored a publication that received an ICLR Spotlight (Top 5% among submissions)

Software Engineer; Edovo; Chicago, IL June 2020 – May 2021

- Led the re-architecture and deployment of a scalable educational platform, handling the full lifecycle from leading product requirement sessions to engineering a data backend
- Architected, tested, and deployed a content platform using Elasticsearch to handle 700K+ requests per day to thousands of users
- Led 10+ requirement gathering sessions with Product owners to re-build a platform
- Created a data pipeline and job to merge 4B rows of user event data in PostgreSQL

Lead Software Engineer; 8th Light; Chicago, IL Jan 2017 – Mar 2019

- Built high-performance technical systems from inception to production, including load testing platforms, API integrations, and patient-facing healthcare applications
- Implemented and deployed a scalable load testing platform simulating 1000+ RPS
- Deployed API integrations to sync 1000+ interactions/ minute in different timezones
- Developed a diabetes management iOS app to connect 60+ patients with diabetic nurses
- Mentored peers and residents through pair programming sessions and code reviews