

Vinay Kudari

Interested in systems that shape user behavior at scale

vinay.kudari30@gmail.com

+1-469-943-6778

RELEVANT SKILLS

Languages: Python, C++, SQL

ML & Deep Learning: PyTorch, JAX, Hugging Face, PySpark

GenAI & Agents: LangGraph, vLLM, QLoRA, RAG

Ranking & Search: ANN, Embeddings, Contrastive learning

Infra & Backend: Kubernetes, FastAPI, Redis, Snowflake, GCP, AWS

EXPERIENCE

- **Apple** Cupertino, CA
Machine Learning Engineer | ***fine-tuning, inference, agents*** *July 2024 - Present*
 - Engineered an enterprise NL-to-SQL pipeline over 10K+ warehouse tables; fine-tuned a 7B planner (Qwen2.5) via FSDP on 100K+ samples, utilizing RAG with negative rejection to ground JSON query plans
 - Delivered a policy-aware invoice intelligence system that parses receipts, enriches with Snowflake policy data, and runs evidence-backed validations to automate “read–query–verify–sign” for 5K+ invoices/day
 - Designed a deep-research finance analyst agent that turns natural language into dashboards and slide decks—retrieving from databases/flat files to generate evidence-backed insights, visualizations, and cited narratives with grounding
 - Engineered an LLM-driven, multi-agent CI/CD pipeline that automates developer onboarding by dynamically reviewing uploaded codebases for syntax and security vulnerabilities before autonomously merging and creating pull requests.
- **Gap Inc** San Francisco, CA
Software Engineer | ***ranking, retrieval, e-commerce*** *July 2023 - June 2024*
 - Engineered a real-time Click-Through Rate (CTR) prediction model using XGBoost to dynamically rank and serve personalized credit card advertisements; integrated user behavioral features, coupon usage, and credit eligibility proxies to optimize ad placement.
 - Engineering a low-latency LLM-based information retrieval system to extract data from multi-source document corpus
- **Amazon Web Services** Palo Alto, CA
Software Development Engineer | ***big-data, etl, pyspark, athena, dynamodb*** *Dec 2022 - June 2023*
 - Designed and published an internal query builder python package (ZenQL) tailored for AWS Healthlake
 - Designed an event-driven server-less system using AWS services, allowing for parallel execution of tasks, resulting in seamless scaling to handle concurrent requests and improved processing throughput
 - Designed and optimized complex SQL queries in Athena resulting in a 40% reduction in query execution times
 - Architected near real time data processing system using AWS CDK, leveraging PySpark for ETL transforms on multi-terabyte healthcare data sets
- **Brave Orbit (early stage startup)** Capetown, South Africa
Founding engineer | ***mobile-app, web-app, django, flutter*** *Jan 2020 - Aug 2021*
 - Programmed a cross-platform app using BLoC pattern, implemented database, network and state management layers
 - Engineered HIPAA compliant scalable async push notification service for medication reminder application
 - Developed a task scheduler for clinical surveys that delivers email, SMS alerts and gathers response on a recurring basis
 - Architected a custom ERP system with multi-product substitutions and combo-product life cycle tracking features

PROJECTS

- **Pixie:** Built in 6 hours at the Nano Banana Hackathon; enables selective edits on live website screenshots or mockups using natural language or voice, with multi-model generation and real-time canvas updates
- **Friend (@mydawg_bot on telegram):** Building a multi-agent AI companion with human-like memory simulation, capable of personalized conversations, context-aware recall, and tool use
- **Stocklerts:** Built a self-evolving system that analyzes real-time news to recommend stocks for day trading; adapts prompts weekly based on ground-truth top performers to refine future predictions
- **Open-source contributions:** Contributor to HuggingFace Datasets official repository

EDUCATION

- **University at Buffalo** Buffalo, NY
Master of Science in Computer Science (AI/ML Specialization)
Research: Finding inconsistency between text and related images, advised by Prof. David Doermann
- **National Institute of Technology, Durgapur** West Bengal, India
Bachelor of Technology in Computer Science and Engineering