Saket Garodia

AI/ML Engineer | AI/ML, RAG, MLOps, Agentic AI

Open to Relocation (SF Bay Area, Seattle, NYC, Remote) | saketgarodia1@gmail.com | 513-807-5167

saketgarodia.com | linkedin.com/in/saket-garodia | github.com/garodisk | medium.com/@SaketGarodia

Summary

AI/ML Engineer with 8+ years across machine learning, MLOps, optimization, NLP, and production AI systems, built on a full-stack software engineering foundation. At Kroger, built and scaled production pricing and promotion optimization systems generating millions of item-location recommendations – **$40M+** in annual incremental profit – and led eCommerce lookalike models scoring **35M+ households** with **$20M+** in potential value. Hands-on GenAI builder with practical experience shipping agentic AI applications, RAG systems, and full-stack LLM products end to end; familiar with fine-tuning, model compression, and alignment workflows through applied projects using BERT, QLoRA, SFT, and DPO. Core stack: PyTorch, LangGraph, OpenAI Agents SDK, Databricks, MLflow, FastAPI, PySpark, SQL, AWS, GCP, and Azure.

Experience

Senior Data Scientist – Machine Learning / MLOps Jan 2022 – Present

84.51° (Kroger, world's largest grocery retailer) *Chicago, IL*

- Built and scaled production pricing and promotion optimization systems generating millions of item-location recommendations across enterprise retail workflows – **$40M+** in annual incremental profit – via forecasting models, Pyomo-based solvers, large-scale data pipelines, and cross-repository CI/CD.
- Led end-to-end development of pickup, delivery, and eCommerce lookalike/lookahead models (**0.84 AUC**) scoring **35M+ households** and supporting **$20M+** in potential value; deployed via Databricks/MLflow with drift monitoring, CI/CD, production testing, and A/B measurement connecting model performance to business impact.
- Re-architected digital engagement segmentation using massive clickstream data across **28M+ households**, replacing a legacy rules-based pillar system now used by Kroger targeting and marketing teams to personalize digital coupon campaigns.
- Built and scaled **RPO-Upkeep**, a rules-based repricing engine automating optimization triggers from cost and competitor price changes; translated guardrails such as size parity, brand spread, cost constraints, and competitive reactions into production logic with **$5M+** potential annual value.
- Won a **2025 84.51° / Kroger Innovation Days hackathon** with **Forecast API Copilot**, a multi-agent GenAI system using OpenAI Agents SDK, GPT-4o, RAG over ChromaDB, Pydantic validation, internal forecasting APIs, and Gradio for natural-language pricing scenario analysis; **selected for productionization evaluation**.
- Sustained **90%+ SLA** across millions of monthly optimization runs with engineering partners; mentored junior data scientists and translated ambiguous retail problems into ML solutions with product, business, and executive stakeholders.

Data Scientist – Machine Learning / NLP Oct 2020 – Dec 2021

Asurion *Remote*

- Reduced Home+ product churn by **20%+** with a churn-risk model built on 100+ features spanning customer attributes, engagement, and POS data, deployed on AWS SageMaker to drive personalized retention strategies.
- Boosted call-center messaging sales per 100 contacts by **15%** by fine-tuning and deploying a BERT-based sentiment model on call transcripts – real-time agent recommendations gated upsell offers on negative-polarity calls, protecting revenue and reducing churn.
- Built NLP capabilities across call-center workflows – keyword detection, topic mining, call routing, and live agent reply suggestions – in cross-functional collaboration with operations, product, and engineering.

Associate, Machine Learning / Data Science Apr 2018 – Aug 2019

Edelweiss Financial Services *Mumbai, India*

- Reduced loan defaults by **30%+** with a credit-risk model (Gradient Boosting, Random Forest; **0.85 AUC**) using demographic, loan-purpose, and geospatial features to predict delinquency probability.
- Developed customer profiling and segmentation frameworks (PCA, K-Means, DBSCAN) for targeted marketing and delivered Tableau dashboards to senior stakeholders.

Application Engineer, Full Stack Jul 2014 – Aug 2015

Oracle *Bangalore, India*

- Built features for Oracle CRM Cloud as a full-stack engineer – data model design, frontend development, and service-layer integrations – and resolved Jira-driven bug fixes and enhancements.

AI/ML Skills

LLMs & Agentic AI: OpenAI Agents SDK, LangGraph, LangChain, CrewAI, MCP, Multi-Agent Systems, RAG, Function Calling, Prompt Engineering, Structured Outputs (Pydantic), Vector Databases (Qdrant, ChromaDB), Embeddings, LLM Evaluation, RAG Evaluation

Fine-Tuning & Alignment: SFT, QLoRA, DPO, PEFT, Quantization (FP16/INT4 NF4), Knowledge Distillation, Model Compression, Hugging Face Transformers

ML & Deep Learning: PyTorch, TensorFlow, Scikit-learn, XGBoost, LightGBM, Random Forest, Gradient Boosting, BERT/DistilBERT, Regression, Classification, Clustering, NLP, Time-Series Forecasting, Mathematical Optimization (Pyomo), A/B Testing, Causal Inference

MLOps / Production AI: Databricks, MLflow, Model Deployment & Serving, Batch & Real-Time Inference, REST APIs, Drift & Performance Monitoring, LLM Observability & Guardrails, CI/CD, GitHub Actions, Docker, Terraform, FastAPI, Prefect, AWS SageMaker, GCP Cloud Run, Azure

Data Engineering: SQL, PySpark, Delta Lake, ETL Pipelines, Data Validation, Tableau

AI Projects & Open Source

Coder Buddy | *LangGraph, GPT-4o, Claude, Gemini, Groq – PyPI: coder-buddy*

- Open-sourced a Claude Code–style agentic coding assistant that converts natural language into working codebases via a LangGraph-orchestrated 4-agent pipeline (Clarifier → Planner → Architect → Coder) with human-in-the-loop gates.
- Built a sandboxed tool layer (file I/O, glob, regex search, shell execution with dangerous-command blocking) with multi-provider LLM support; published to PyPI.

AI Consultation Assistant – Production GenAI SaaS | *Next.js, FastAPI, OpenAI, Clerk, Docker, AWS*

- Built and deployed a full-stack GenAI SaaS that converts consultation notes into structured summaries, action items, and client-ready email drafts – real-time SSE streaming, structured LLM outputs, Clerk authentication, and subscription-gated access.
- Containerized with Docker and deployed to AWS App Runner via ECR with CI/CD, environment management, and logging, behind a clean Next.js/FastAPI REST architecture.

RAG Search System | *Next.js, FastAPI, Qdrant, BGE, BM25/RRF, Prefect, GCP*

- Built and deployed a full-stack hybrid-search RAG system (BGE-base embeddings + BM25 with reciprocal-rank fusion) with streaming AI Q&A over 50K+ Medium articles – Next.js frontend on Vercel, FastAPI backend on GCP Cloud Run, Qdrant vector DB.
- Automated MLOps pipeline (Prefect + FastEmbed) handles continuous RSS ingestion and incremental embedding updates.

IT Ticket Classifier – BERT Fine-Tuning & Compression | *PyTorch, Hugging Face, DistilBERT*

- Fine-tuned BERT-base on 47.8K IT tickets (**88.2%** accuracy, 87.9% macro-F1), then compressed via knowledge distillation into DistilBERT and FP16/INT4 (NF4) quantization – **74%** memory reduction (255 MB → 66 MB) with $<$0.1% accuracy loss; published models and dataset to Hugging Face Hub.

LLM Alignment Pipeline | *QLoRA, SFT, DPO, Llama-2-7B, Mistral-7B*

- Applied QLoRA-based SFT to Llama-2-7B (**71%** perplexity reduction with 40M trainable parameters of 6.8B), then aligned Mistral-7B with DPO on 12.8K preference pairs (final loss 0.040) without a reward model; published to Hugging Face.

Build & Learn: GPT from Scratch | *PyTorch – github.com/garodisk/GPT-from-scratch*

- Implemented a character-level GPT from scratch in PyTorch – multi-head self-attention, LayerNorm, residual connections, and full training loop – published as a 4-part Medium series with companion GitHub repo.

Education

M.S. Business Analytics, University of Cincinnati, OH – GPA 4.0/4.0, Graduate Merit Scholarship — *Aug 2019 – Aug 2020*

B.Tech, Computer Science & Engineering, National Institute of Technology (NIT), Calicut, India — *Jul 2010 – Jun 2014*

MBA, International Business, Indian Institute of Foreign Trade (IIFT), New Delhi, India — *Jul 2016 – Mar 2018*