Mihir D. Chauhan

mihirchauhan951@gmail.com | (720) 813-5491 | LinkedIn | GitHub | Portfolio

Education

University of Colorado Boulder — Boulder, CO
M.S. in Data Science, GPA: 3.96/4.00 — *Aug 2024 – May 2026*
Relevant Coursework: Machine Learning, Natural Language Processing, Statistical Methods, Data Mining, Data Scale Systems

University of Mumbai — Mumbai, India
B.E. in Computer Engineering, GPA: 9.29/10.00 — *Aug 2020 – May 2024*

Skills

Languages: Python, SQL, JavaScript, TypeScript, R, C++

AI / ML & Backend: LLMs, RAG pipelines, multi-agent orchestration, LangChain, LangGraph, prompt engineering, PyTorch, TensorFlow, scikit-learn, BERT, MLOps, human-in-the-loop learning, FastAPI, async Python, PostgreSQL, Redis, Supabase, OpenAI / Gemini / Anthropic APIs

Cloud & DevOps: AWS (ECS Fargate, RDS, S3, CloudWatch, Secrets Manager), Docker, Terraform, GitHub Actions, CI/CD, Nginx, PM2

Experience

University of Colorado Anschutz — Aurora, CO
Research Intern — Clinical AI — *Jan 2026 – Present*

- Surfaced systematic LLM failure modes and clinical relevance gaps across healthcare annotation workflows by building rubric-based QA pipelines and automated error analysis tooling, improving model monitoring coverage and output reliability.
- Accelerated clinical AI artifact quality by running systematic prompt engineering and structured-output experiments, iterating on generation templates, safety guardrails, and output schemas to increase consistency and domain specificity across diverse healthcare use cases.

Honda 99P Labs — Boulder, CO
AI Engineer | Blog — *Sep 2025 – Dec 2025*

- Architected a multi-agent debate system end-to-end with hybrid retrieval, vector databases, and curated knowledge packs; built modular API abstractions decoupling the dialog engine from retrieval, scoring, and response generation to enable reproducible, context-aware AI reasoning.
- Delivered a FastAPI backend with advanced RAG, context engineering, and AWS CloudWatch observability; deployed on ECS Fargate with Secrets Manager integration for secure, scalable cloud delivery with end-to-end structured logging throughout.

Kobeyo — Boulder, CO
Junior Data Scientist Intern — *May 2025 – Aug 2025*

- Improved production classification accuracy by building an MLOps pipeline for a custom BERT model with human-in-the-loop feedback loops, enabling continuous fine-tuning and structured retraining cadences that sustained accuracy gains in a live environment.
- Cut query time by 50% and expanded business data coverage by 40% by building a full-stack Python ingestion API with async Playwright scraping and Supabase (PostgreSQL) backend; applied LLMs with prompt engineering to automate skill-tagging and structured job data extraction at scale.

Projects

AgentSquared — No-Code AI Agent Platform — Boulder, CO
Live Demo | HackCU 2026 — *Mar 2026*

- Built a full-stack no-code AI agent platform in 24 hours (FastAPI, Next.js, Gemini API) enabling small businesses to deploy trained agents in under 60 seconds; architected 3 distinct agent types across 15+ REST endpoints and 8 frontend routes over a shared RAG backbone, shipping 4,000+ lines of production code fully deployed on Vultr.
- Engineered an async RAG pipeline with website crawling and PDF ingestion to dynamically ground agent responses in proprietary business knowledge; integrated Bluesky AT Protocol with Gemini-powered sentiment analysis and auto-reply generation; deployed with Nginx and PM2 for production-grade reliability under live traffic.

CodeSense — AI-Powered Code Review Platform — Boulder, CO
Independent Project — *Oct 2025 – Dec 2025*

- Architected a scalable multi-tenant FastAPI backend with 15+ REST endpoints, JWT authentication, and async Redis RQ workers; integrated GitHub webhooks with GPT-powered patch generation and automated static analysis (Semgrep, Ruff, Bandit) to deliver autonomous, real-time code review across concurrent tenants.
- Provisioned full AWS cloud infrastructure via Terraform (ECS Fargate, RDS, ElastiCache, S3, CloudFront, CloudWatch) within a secure VPC; managed end-to-end containerization and CI/CD through Docker and GitHub Actions for reliable, reproducible deployments.

Yoga Posture Correction and Detection — Mumbai, India
Team Lead | Published Paper — *Aug 2023 – May 2024*

- Published peer-reviewed research and led a team of 4 to develop a TensorFlow-based real-time pose estimation model achieving 96.5% accuracy across 12 poses; processed 3,600+ images, deployed via TensorFlow.js for live inference, and built an interactive platform driving 40% higher user participation.