Sahil Mehta

New York City, NY · Open to relocation | sahilmehta0204@gmail.com | 608-960-5508 | sahilmehta.dev | Github | Linkedin

EDUCATION

University of Wisconsin, Madison
B.S. in Computer Science | B.S. in Data Science (double major) *Graduation: May 2025*

PROFESSIONAL EXPERIENCE

AI/Full-Stack Engineer, Enidus USA LLC. (Full-Time) *June 2025 - Present* | **Hicksville, NY**

AI Chatbot & Agentic Copilot for T-Mobile for Business

- Shipped a multi-tenant conversational AI assistant for a T-Mobile IoT reseller platform: account admins query telecom account data and run multi-step transaction flows (device orders, line state, plan changes) in natural language; in pilot with 15 tenants, 25+ customers, 100+ users.
- Designed a hybrid retrieval pipeline combining Qdrant vector search over per-tenant collections with BM25 keyword retrieval and Reciprocal Rank Fusion (k=60), plus an optional Cohere/Voyage re-ranker; deployed on AWS.
- Architected a vector-first routing inversion when the product pivoted consumer-facing; calibrated the confidence threshold against a 442-query eval corpus, moving intent top-1 accuracy from 73.5% to 89.0%.
- Built defense-in-depth safety: tool-gating at the LLM boundary (43 Pydantic-typed tool handlers, RBAC-filtered catalog the model never sees in full), parameterized SQL templates with row-level security at execution, per-tenant Qdrant isolation, and user-confirmed writes, with zero hallucination incidents in pilot.
- Cost- and latency-engineered for consumer scale: a Qdrant result cache and embedding cache, regex fast-paths that bypass embedding for identifier inputs, and confidence-bucketed routing that skips LLM round-trips on high-confidence intents; routing and response regressions surface to telemetry dashboards.

Custom Reports & Dashboards Platform

- Owned a self-serve full-stack analytics product end-to-end alone (React, Node.js + Express, SQL Server with stored procedures) letting enterprise customers compose reports, custom dashboards, and charts over their own data; cut analytics turnaround from days to minutes.
- Hardened against multi-tenant attack classes via stored-procedure CRUD contracts, two-layer filter validation, runtime tenant-clause injection, JWT auth, per-session CSRF rotation, strict CSP, and AES-256-CBC encryption for stored Excel passwords.

Software Developer, Orahi (Internship) *July 2024 - August 2024* | **Remote**

- Designed a dynamic bus route adjustment algorithm using K-means clustering, reducing manual student-assignment effort by 80%; optimized Flask REST APIs for telemetry ingestion.

Data Scientist, GSPANN Technologies Inc. (Internship) *June 2023 - August 2023* | **Remote**

- Built a CNN-based pneumonia detection model on chest X-ray images; iterated on preprocessing and data augmentation to improve generalization.

PROJECTS

ClaudeJob - Agentic Resume Tailoring Pipeline (Github) *April 2026 - Present | Personal Project*

- Built an end-to-end agentic pipeline (Node.js + Anthropic SDK + SSE streaming) that ingests live job postings, tailors a structured-output JSON resume per role, and generates pixel-matching PDFs via pdfkit; actively used to power my own AI Engineer applications.
- Engineered a validator suite mirroring LLM-content failure modes: 30+ banned AI-resume cliché regex, source-fact validation against a pinned base to catch fabricated stats, and a jargon-lead heuristic. Deterministic adjacency-skill injection (curated, never LLM-fabricated); 47 passing unit tests.

RAG Pipeline - Denari AI Capstone *January 2025 - May 2025 | Madison, WI*

- Productionized a full-stack RAG system over 22K+ documents and 300K+ embeddings (TypeScript, TimescaleDB, Docker, S3, OpenAI APIs); hybrid retrieval (BM25 + TF-IDF) with semantic re-ranking achieving 73% QA accuracy and 40% query-latency reduction; led Agile/Scrum delivery of 25+ production features across ingestion, embeddings, DB, and retrieval.

TECHNICAL SKILLS

AI / LLM Systems: LLM APIs (Claude, OpenAI), tool calling, agent orchestration, RAG, vector search (Qdrant), prompt engineering, eval frameworks, structured outputs (Pydantic), streaming/SSE, sentence-transformers, lifecycle marketing, PyTorch, TensorFlow
Languages: Python, JavaScript/TypeScript, Java, C, SQL, Kotlin, Swift, R. Cert: SnowPro Associate & Core (2024).
Frameworks: FastAPI, Node.js, Express, React, Next.js, Angular, Flask, Django, React Native
Infra & Tools: PostgreSQL, REST, gRPC, AWS S3, GCP, Docker, Kubernetes, Git, Claude Code, Claude Code Skills, MCP, Braze, JWT/OAuth, RBAC