Kunj Rathod

Software Engineer Intern — AI, Backend, and Cloud Systems

+1 (385) 202-8879 — kunj.rathod@utah.edu — linkedin.com/in/rathodkunj — github.com/rathodkunj2005

Education

Bachelor of Science in Computer Science Aug 2023 – Dec 2026
University of Utah, Salt Lake City, UT *GPA: 3.6/4.0*

Technical Skills

- **Languages:** Python, Java, C++, TypeScript, JavaScript, SQL
- **Frameworks and Tools:** React, Flask, Node.js, SwiftUI, Docker, CI/CD, Git, Vite
- **Databases:** PostgreSQL, DynamoDB, MySQL, Vector databases (Pinecone, Chroma), NebulaGraph
- **Cloud and Infrastructure:** AWS (Lambda, S3, Bedrock, API Gateway, RDS), API Design, Microservices
- **AI/ML:** RAG/GraphRAG, LangChain, LlamaIndex, OpenAI API, Model Deployment, LLM Optimization, TensorFlow

Experience

Software Development Intern, AI Services Jan 2025 – Present
University of Utah SUDO Program, Salt Lake City, UT

- Built and deployed a HIPAA-compliant AI chat platform for 90+ hospital executives (React/TypeScript, Flask, AWS Bedrock microservices, event-driven Lambda orchestration).
- Shipped 6 full-stack features across 4 sprints; integrated Bedrock Agents/Knowledge Bases/Guardrails for production workflows; owned API design, schema, UI components, and AWS infrastructure deployment.
- Implemented token-streaming LLM responses (p95 < 200ms time-to-first-token) with resilient fallback handling and distributed session persistence (DynamoDB ephemeral state + S3 durable storage) for 1,000+ conversations, ensuring HIPAA compliance.

Undergraduate Researcher, Agentic Ashby Plot Generation for Materials Discovery Aug 2025 – Jan 2026
STARS Lab, University of Utah

- Built a multi-agent, graph-augmented pipeline to extract and normalize material-property data (tables/figures) from 1,000+ materials-science papers into a physics-aware graph for automated Ashby plot generation.
- Developed a constraint-based “design region” engine (e.g., temperature/creep/pressure limits) and benchmarking suite (extraction accuracy, plot fidelity) to identify feasible materials for extreme environments.

AI Engineering Intern Nov 2024 – Apr 2025
CourtEasy.ai (Remote)

- Scaled hybrid legal-document retrieval to 10M+ indexed documents, supporting 5,000+ daily queries for an AI legal research platform.
- Improved retrieval accuracy by 28% and reduced hallucinations by 35% by implementing hybrid RAG (dense vectors + BM25 + reranking) and context-grounding optimizations.
- Built production ETL ingesting 500k+ documents/week (normalization, entity extraction, quality gates) and benchmarked 8 LLM families on LegalBench to guide model routing and reduce projected inference spend ($50k+/yr).

Technical Projects

Wingman.ai — Multi-Modal AI Personal Assistant

SwiftUI, OpenAI API, Firebase, MVVM Apr 2025 – Present

- Created an iOS personal assistant with voice, chat, and image input modes, integrating GPT4.1 and Whisper APIs for context-aware responses.
- Implemented offline-first architecture with Firebase sync supporting real-time message streaming and conversation history.

Minute0 – AI-Powered Deployment Monitor *Feb 2026*
React, TypeScript, Cerebras, FastAPI, ChromaDB, Slack API minute0.vercel.app

- Built a full-stack deployment monitoring + incident response system that tracks Vercel deployments, classifies build/runtime failures, and opens incidents with Slack alerts and approval workflows.
- Implemented AI-assisted root-cause analysis with FastAPI and ChromaDB vector search over logs/errors, generating structured fix suggestions for downstream coding agents.
- Delivered a real-time React/TypeScript dashboard for live metrics, incident status, and agent health; deployed on Vercel with CI/CD.

BioGraphRAG — Biomedical Knowledge Graph Retrieval System

Python, NebulaGraph, LlamaIndex, Docker, AWS May 2024 – Jan 2025

- Engineered distributed GraphRAG system managing 1M+ biomedical entities (proteins, genes, diseases) with graph-augmented LLM retrieval, improving factual accuracy in biomedical Q&A by 40%.
- Optimized graph traversal performance 3x through strategic caching and high-degree node pruning, supporting under 500ms query latency at p95.
- Designed automated ETL pipelines ingesting UniProt and AlphaFold datasets, processing 2M+ entity updates monthly with schema validation.