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, Computer Science (Expected Dec 2026) Aug 2023 – Dec 2026
University of Utah, Salt Lake City, UT *GPA: 3.6/4.0*
Relevant Coursework: Machine Learning, Computer Vision, NLP, Data Structures and Algorithms, Database Systems

Technical Skills

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

Experience

Software Development Intern, AI Services Jan 2025 – Present
University of Utah SUDO Program, Salt Lake City, UT
- Built and deployed HIPAA-compliant AI chat platform serving 90+ hospital executives using React/TypeScript frontend, Flask backend, and AWS Bedrock microservices with event-driven Lambda orchestration
- Shipped 6 full-stack features across 4 sprint cycles in fast-paced environment, integrating AWS Bedrock Agents, Knowledge Bases, and Guardrails for production AI workflows
- Implemented streaming LLM integration with token-by-token rendering achieving under 200ms time-to-first-token (p95) with robust fallback handling for network failures
- Architected distributed session persistence managing 1,000+ chat conversations with ephemeral state in DynamoDB and durable storage in S3, ensuring HIPAA compliance
- Owned end-to-end development of multiple full-stack applications, from API design and database schema to frontend components and AWS infrastructure deployment

AI Engineering Intern Nov 2024 – Apr 2025
CourtEasy.ai (Remote)
- Scaled legal document retrieval system to 10M+ indexed documents, serving 5,000+ daily queries for AI-powered legal research platform
- Improved retrieval accuracy by 28% and reduced hallucination rate by 35% through custom RAG pipeline combining dense vector search, BM25 reranking, and context-window optimization
- Built production ETL pipeline processing 500,000+ documents weekly with text normalization, entity extraction, and quality filtering to maintain corpus integrity
- Benchmarked 8 LLM families (GPT-4, Claude, Llama, Mistral) on LegalBench, optimizing inference budget ($50,000+/yr) and guiding multi-model pipeline design

Undergraduate Researcher, AI for Materials Science Aug 2025 – Present
STARS Lab, University of Utah
- Building multi-agent GraphRAG system integrating domain-specific retrieval and structured reasoning over 100,000+ materials science papers, enabling automated Ashby plot generation for aerospace alloy analysis
- Developed computer vision pipeline using OpenCV and OCR to extract mechanical property data from 30+ research figures with 92% parsing accuracy, automating literature-to-database ingestion for high-temperature alloys
- Integrated tool-calling architecture connecting LLM reasoning modules with symbolic math (SymPy) and finite element simulators, improving material property prediction accuracy by 18%

Technical Projects

BioGraphRAG — Biomedical Knowledge Graph Retrieval System
Python, NebulaGraph, LlamaIndex, Docker, AWS *May 2024 – Jan 2025*
- Built 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

Wingman.ai — Multi-Modal AI Personal Assistant (Independent Project)
SwiftUI, OpenAI API, Firebase, MVVM *Apr 2025 – Present*
- Built iOS personal assistant with voice, chat, and image input modes, integrating GPT-4V and Whisper APIs for context-aware responses
- Implemented offline-first architecture with Firebase sync supporting real-time message streaming and conversation history across devices for 100+ beta users