Kunj Rathod

Computer Science Researcher & AI Engineer

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

Summary

AI/Software Engineer specializing in LLM systems, Retrieval-Augmented Generation (RAG), scalable cloud infrastructure, and embodied AI research. I design and ship AI applications across healthcare, legal-tech, aerospace, and robotics — from HIPAA-compliant hospital platforms to spatial memory systems for embodied agents, distributed biomedical knowledge graphs, and multi-agent materials discovery pipelines. Currently an incoming Software Engineering Intern on the **Microsoft Azure Data** team and an ongoing AI Services Intern at **University of Utah Health**.

Education

Bachelor of Science in Computer Science
Aug 2023 – Dec 2026
University of Utah, Salt Lake City, UT
GPA: 3.7/4.0 | Dean's List

- **Relevant Coursework:** Machine Learning, Computer Vision, Natural Language Processing, Distributed Systems, Algorithms & Data Structures, Interpretability of LLMs, Multimodal LLM Agents

High School Diploma
2017 – 2023
Krishna Public School, Raipur, India

Professional Experience

Software Engineer Intern — Azure Data
May 2026 – Present
Microsoft, Redmond, WA

- Joining the Azure Data team for Summer 2026 to build scalable cloud solutions for distributed data systems.
- Focusing on full-stack software development and distributed systems within the Azure ecosystem.

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

- Built and deployed a HIPAA-compliant AI chat platform for 90+ hospital executives using React/TypeScript frontend, Flask middleware, and AWS Bedrock microservices with event-driven Lambda orchestration.
- Shipped 6 full-stack features across 4 sprints; integrated AWS Bedrock Agents, Knowledge Bases, and Guardrails for production clinical workflows; owned API design, schema, UI components, and infrastructure deployment.
- Reduced inference latency by 40% and data query speed by 60% through AWS Bedrock pipeline optimization, API caching, and a DynamoDB–RDS hybrid database strategy.
- 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.
- Integrated interactive data visualization tools into the LLM chat interface enabling real-time analytics on hospital data within conversational flows.

Undergraduate Researcher — LLMs & Computational Simulations
Aug 2025 – Feb 2026
STARS Lab, University of Utah | Collaboration: NASA, Microsoft, U.S. DoD
Salt Lake City, UT

- Built a multi-agent, graph-augmented pipeline to extract and normalize material-property data (tables and figures) from 1,000+ materials-science papers into a physics-aware graph for automated Ashby plot generation.
- Developed a constraint-based "design region" engine (temperature, creep, and pressure limits) and benchmarking suite (extraction accuracy, plot fidelity) to identify feasible materials for extreme aerospace environments.
- Explored applications of LLMs and multi-agent AI to streamline knowledge sharing across interdisciplinary stakeholders including engineers, scientists, and DoD partners.
- Built Ref-RAG, a custom RAG chatbot using LangChain and Chainlit to extract structured information from large unorganized PDF datasets for materials researchers.
- Contributed to high-throughput experimentation, computational modeling, and AI-driven materials design strategies for rocket engines and hypersonics.

AI Engineering Intern
Nov 2024 – Apr 2025
CourtEasy.ai / Nugen
Remote

- Scaled hybrid legal-document retrieval to 10M+ indexed Indian legal documents (statutes, court orders), 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 for Legal-NER tasks.

- Built production ETL ingesting 500k+ documents/week (normalization, entity extraction, quality gates) and benchmarked 8 LLM families on 4 legal benchmarks including LegalBench and NyayaAnumana.
- Evaluated InLegalBERT, InLegalLLaMA, and GPT-4o-mini on F1 score, latency, and token-level cost metrics; analysis guided model routing decisions reducing projected inference spend by $50k+/yr.
- Co-authored a comparative analysis paper synthesizing insights from 15+ research papers on legal AI, informing the team's LegalBench-RAG workflows and evaluation protocols.

Campus Strategist *Jan 2025 – Apr 2025*
Perplexity AI Salt Lake City, UT

- Spearheaded campus-wide outreach programs to accelerate adoption of Perplexity's AI-powered search platform among students, faculty, and university clubs.
- Onboarded 150+ Perplexity Pro users, facilitating seamless onboarding and sustaining long-term engagement across campus communities.

AI Research Intern — BioGraphRAG | GMG Summer of Code *May 2024 – Aug 2024*
Garje Marathi Global Salt Lake City, UT

- Led development of BioGraphRAG, a Graph Retrieval-Augmented Generation platform combining biomedical knowledge graphs with LLMs for accurate, explainable answers to complex biomedical queries.
- Engineered distributed GraphRAG system managing 1M+ biomedical entities (proteins, genes, diseases) integrating UniProt, AlphaFold, and RXNav datasets with NebulaGraph for storage and indexing.
- Improved factual accuracy in biomedical Q&A by 40%; optimized graph traversal performance 3× through strategic caching and high-degree node pruning, achieving sub-500ms query latency at p95.
- Designed automated ETL pipelines processing 2M+ entity updates monthly with schema validation using Python, Docker, LlamaIndex, and FastAPI.
- Conducted node degree analysis to identify and eliminate root causes of high-latency responses, significantly improving system performance.
- Presented BioGraphRAG at an AI panel discussion attended by experts from India and the US, receiving commendation for technical leadership and team mentorship.

Research

Spatial Memory for Embodied Agents & Long-Horizon Web Agents *2026 – Present*
University of Utah

- Investigating long-horizon task solving for web agents, studying Web Explorer and Web Sailor V2 for extended agentic reasoning and trajectory-based multi-turn interaction.
- Evaluating agent benchmarks including OS Marathon (Feb 2026), BrowseComp, and Mind2Web across live and offline evaluation settings to assess real-world web task performance.
- Researching retrieval-augmented spatial memory architectures for embodied agents (inspired by ReMEmbR / NaVQA), unifying spatial, temporal, episodic, and semantic memory for long-horizon robot navigation and manipulation.
- Studying agentic scene generation (SceneSmith, MIT CSAIL / Toyota Research) for automatically constructing simulation-ready environments from natural language, enabling scalable, automatic robot policy evaluation.

Predicting Generalization from Circuits using LLM Analysis for Interpretability of LLMs *2026 – Present*
University of Utah

- Researching circuit-level generalization in LLMs using sparse feature circuits, based on Wu et al. on LLM performance on non-default vs. default tasks.
- Building automated LLM pipelines to extract circuits from default tasks and predict generalization to non-default tasks; using attribution graphs and Pathways Discovery (PD) for circuit-level analysis.
- Responsible for citations analysis and synthesizing prior work on circuit stability and generalization correlation.

Agentic Ashby Plot Generation for Aerospace Materials Discovery *Aug 2025 – Feb 2026*
STARS Lab, University of Utah

- Developed an end-to-end multi-agent pipeline for automated materials selection using graph-augmented retrieval over scientific literature.
- Created benchmarking suites for extraction accuracy and plot fidelity, enabling systematic evaluation of AI-driven materials discovery approaches.

BioGraphRAG: Biomedical Knowledge Graph Retrieval System *May 2024 – Jan 2025*
Garje Marathi Global / GMG Summer of Code

- Designed and implemented a GraphRAG architecture integrating heterogeneous biomedical databases (UniProt, AlphaFold, RXNav, BioKG) into a unified NebulaGraph knowledge store.
- Investigated graph-structured retrieval as a mechanism to reduce hallucinations and context inaccuracies in LLM-

generated biomedical answers.

Technical Projects

FNDR — Privacy-First Local AI Assistant for macOS *Jan 2024 – Present*
Rust, Tauri, Metal, ONNX, Llama 3.2, Whisper, Apple Vision Framework

- Engineered a high-performance macOS desktop application using **Rust and Tauri**, delivering a zero-trust, local-only memory assistant with full data sovereignty — no cloud, no telemetry.
- Optimized on-device inference for LLMs (Llama 3.2) and VLMs (SmolVLM) with **Metal-accelerated backends**, achieving low-latency RAG on M-series Apple Silicon.
- Architected a real-time screen extraction pipeline using **Apple Vision Framework** for high-speed OCR and **CLIP-based visual embeddings** to reconstruct temporal context from screen snapshots.
- Designed a **Graphiti-style Temporal Search Engine** modeling semantic relationships across user activities, web sessions, and meeting transcripts, enabling proactive entity extraction and multi-hop reasoning.
- Implemented automated meeting intelligence with local **Whisper-based transcription** (Parakeet) and segmented audio processing integrated into the global memory index.
- Developed a **Model Context Protocol (MCP)** server for secure, local interoperability between the memory store and external AI agents or IDEs.

HirePilot — Autonomous AI Recruiting Agency *2026*
TypeScript, Node.js, Express, PostgreSQL, Anthropic API

- Built a fully autonomous recruiting backend with specialized AI agents (Enrichment, Scheduling, Interview, Evaluation) to manage the end-to-end hiring lifecycle, from GitHub sourcing to live candidate screening.
- Engineered complex integrations with Twilio for real-time voice AI interviews (with transcript analysis), Google Calendar for automated slot scheduling, and Slack/Resend for manager approvals and multichannel outreach.

Minute0 — AI-Powered Deployment Monitor [Hackathon Winner] *Feb 2026*
React, TypeScript, Cerebras, FastAPI, ChromaDB, Slack API | *minute0.vercel.app*

- Built a full-stack deployment monitoring and incident response system tracking Vercel deployments, classifying build/runtime failures, and triggering Slack alerts with approval workflows.
- Implemented AI-assisted root-cause analysis with FastAPI and ChromaDB vector search over logs and 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 pipeline.

Wingman.ai — Multi-Modal AI Personal Assistant *Apr 2025 – Present*
SwiftUI, OpenAI API (GPT-4o, Whisper), Firebase, MVVM

- Created an iOS personal assistant with voice, chat, and image input modes integrating GPT-4o and Whisper APIs for context-aware responses with RAG-enhanced memory.
- Implemented offline-first architecture with Firebase sync supporting real-time message streaming and persistent conversation history.

Ref-RAG — Research Literature Chatbot *Aug 2025 – Feb 2026*
Python, LangChain, Chainlit, FastAPI

- Built a custom RAG chatbot for the STARS Lab to extract structured information from large, unorganized PDF corpora of materials-science research papers.
- Enabled researchers to query domain-specific knowledge across 1,000+ documents through a conversational interface.

FlowVía — V2X Urban Mobility Optimization System *Apr 2024*
Python, TensorFlow (LSTM), V2X (DSRC / C-V2X), OBD-II, Cloud Backend

- Designed a Vehicle-to-Everything (V2X) traffic optimization platform combining V2V, V2I, and V2N communication protocols for real-time adaptive traffic management across urban road networks.
- Implemented real-time speed recommendation algorithms using live SPaT data, and built LSTM-based traffic flow prediction models continuously retrained on historical traffic patterns.
- Architected a full system stack: in-vehicle OBD-II hardware unit, mobile driver interface, cloud ML backend with city traffic API integration, and roadside V2X unit interfaces.
- Designed AES-256 encrypted communication, rotating vehicle identifiers for anonymization, and edge-first processing architecture to meet privacy and ultra-low-latency requirements.

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

- Engineered a production-grade distributed GraphRAG system for healthcare professionals, researchers, and patients requiring trustworthy biomedical information retrieval.

- Integrated UniProt, AlphaFold, RXNav, and BioKG into a unified NebulaGraph store with automated ETL processing 2M+ entity updates monthly.

Technical Skills

Languages	Python, Java, C++, TypeScript, JavaScript, Swift, SQL
Frameworks	React, Flask, Node.js, Next.js, SwiftUI, FastAPI, Chainlit, Docker, Kubernetes, CI/CD, Git, Vite, Tailwind CSS
Databases	PostgreSQL, DynamoDB, MySQL, RDS, Pinecone, ChromaDB, NebulaGraph
Cloud	AWS (Lambda, S3, Bedrock, API Gateway, RDS, Bedrock Agents), Firebase, Microservices, Event-Driven Architecture
AI / ML	RAG, GraphRAG, LangChain, LlamaIndex, CrewAI, OpenAI API, AWS Bedrock, PyTorch, TensorFlow, DistillBERT, DQN, PPO, Transformers, Multi-Agent Systems, Embodied AI, Web Agents
Developer Tools	VS Code, GitHub, Docker, Vercel, Postman

Writing & Publications

BioGraphRAG — Biomedical Knowledge Graph Retrieval Augmented Generation *Oct 2024*
kunjrathod.substack.com/p/biographrag | Kunj's Substack

- Co-authored with Niraj Kumar Singh (ML Engineer). Full technical article presenting BioGraphRAG: system architecture, GraphRAG algorithm, node-degree performance analysis, multi-stage answer enrichment pipeline integrating UniProt, AlphaFold, and RXNav, and future directions. Published as part of the GMG Summer of Code program.
- *Note: NebulaGraph's marketing team reached out requesting republication on their official website (Jun 2025).*

FlowVía: A Technical Deep Dive into Next-Gen Urban Mobility *Apr 2024*
kunjrathod.substack.com/p/flowvia | Kunj's Substack

- Solo-authored technical article covering FlowVía, a V2X urban traffic optimization system. Details system architecture (V2V, V2I, V2N protocols, DSRC and C-V2X standards), real-time speed recommendation algorithms, LSTM-based traffic flow prediction models, data privacy/security design, and scalability challenges.

Comparative Analysis: LLM Families on Legal Benchmarks *2025*
Internal technical report — CourtEasy.ai / Nugen

- Co-authored comparative analysis of InLegalBERT, InLegalLLaMA, and GPT-4o-mini on LegalBench and NyayaAnumana benchmarks, synthesizing insights from 15+ research papers to inform production RAG workflow design and evaluation protocols.

Awards & Recognition

- **Kahlert Impact Prize** — Kahlert School of Computing, University of Utah; $1,000 undergraduate scholarship awarded for societal impact through AI research and production systems in healthcare, legal-tech, and embodied AI, Mar 2026. Funded by a $15M endowment from The Kahlert Foundation; recognizes students with a compelling track record of translating computing research into real-world societal benefit.
- **Minute0 Hackathon Winner** — Deployed full-stack AI deployment monitoring system, Feb 2026
- **Dean's List** — University of Utah, College of Engineering, 2024–2025
- **AI Panel Presentation** — Presented BioGraphRAG at international AI panel attended by experts from India and the US; received commendation for technical leadership, GMG Summer of Code, 2024

Leadership & Involvement

- **Community Advisor** — University of Utah Housing & Residential Education; led safety, conflict resolution, and community programming for 200+ residents (Aug 2024 – Dec 2024)
- **Team Lead** — Code-Crafters Team, GMG–MAAI Summer of Code; led development and cross-functional coordination for BioGraphRAG, mentoring junior contributors (May – Aug 2024)
- **Campus Strategist** — Perplexity AI; led campus outreach and drove 150+ Pro user onboarding across student and faculty communities (Jan – Apr 2025)

Languages

- **English** — Native / Bilingual
- **Hindi** — Native / Bilingual