

Bimal Kumal

917-821-6444 | bkumal@student.sjny.edu | [linkedin.com/in/bimal-kumal](https://www.linkedin.com/in/bimal-kumal) | [Portfolio](#)

EDUCATION

St. Joseph's University, New York,
BS in Computer Science & Data Science
GPA: 3.948

Brooklyn, NY
Expected: May 2027

Relevant Coursework: Data Structures & Algorithms, Advanced Programming, Operating Systems, Database Systems, Linear Algebra, Calculus I, II, Software Engineering, Probability & Statistics

PROFESSIONAL EXPERIENCE

TransRe
Agentic AI Engineer Intern

June 2026 - Present
New York City, NY

- Built a production FinOps AI agent (Claude tool-use loop + custom FastMCP servers) tracking \$600K/month in multi-cloud spend; identified \$150K in one-time infrastructure rightsizing and \$115K/month in recurring savings by reclaiming 400+ inactive M365 licenses.
- Replaced a 3–4-day manual financial reporting process with zero-touch delivery, using Kubernetes CronJobs and a hybrid live-API/Fabric-snapshot pipeline that also eliminated recurring Azure rate-limit failures.
- Architected TransRe's internal AI platform on Azure Kubernetes, replacing per-agent gateways with a single Okta-SSO front door at the Istio edge; cut new-agent onboarding to a one-line database update, with zero code changes required.
- Launched a RAG IT Helpdesk assistant serving ~700 employees, answering questions from the internal knowledge base and auto-escalating unresolved ones to Freshservice; built a hybrid pgvector/full-text retrieval pipeline validated to 91.7% recall@5 on a 222-question evaluation harness.

PROJECTS

Arthvion | AI Due Diligence Platform for Private Equity

July 2026

Tech stack: Python, FastAPI, PostgreSQL, Claude API, MCP, Next.js, Docker, Terraform

- Built a multi-agent research platform (FastAPI, React, PostgreSQL) that pulls SEC filings, market comps, and risk flags into a full investment memo in under 4 minutes — replacing days of manual analysis.
- Cut inference costs by ~90% and accelerated response times 5x using parallel API calls, exponential backoff, and rolling prompt-cache breakpoints.
- Shipped a production-ready architecture: Redis-backed SSE streaming, an atomic SQL ledger for precise billing, rate limiting, and non-root containers.

Pulse | Personal AI Automation Agent [GitHub](#)

April 2026

Tech stack: FastAPI, React, OpenAI, APScheduler, REST APIs, SQLite

- Architected an autonomous GPT-4 agent chaining 11 custom tools across 10+ APIs to auto-rank ~70 job postings/week and generate tailored PDF cover letters — cutting application time from ~30 minutes to under 1 minute and saving ~5 hours/week.
- Deployed 7 daily APScheduler cron pipelines for AI digests and job alerts and secured the public Vercel/Render deployment with OTP auth, server-side token validation, and API rate limiting.

Stellonet | Full-Stack Web Application | [Website](#)

March 2026

Tech Stack: Python, FastAPI, PostgreSQL, AWS Nova, Docker, Vercel, Authentication

- Built a full-stack automated outreach platform (React, Vercel, FastAPI, AWS App Runner) that analyzes student research interests to instantly generate ranked professor matches and personalized communication workflows.
- Integrated Amazon Nova with OpenAlex data under a strict-JSON contract to eliminate hallucinations, reducing match latency by 4x via parallel university fan-out and caching.

SKILLS

Languages: Python, TypeScript, JavaScript, Java, SQL, R

AI & Machine Learning: LangGraph, RAG, Multi-Agent Systems, MCP (Model Context Protocol)

Frameworks & Web: React, Next.js, Tailwind CSS, FastAPI, REST APIs, Server-Sent Events (SSE)

Cloud & DevOps: Azure (AKS, ACR, Key Vault), AWS (EC2, App Runner, Bedrock), Kubernetes, Docker, CI/CD, Git

Databases & Infrastructure: PostgreSQL, pgvector, Redis, Celery, SQLite, Microsoft Fabric