Nikhil Kumar

Hyderabad, India | Open to relocation
nikhilnkup@gmail.com | linkedin.com/in/nkupa | github.com/Nikhil-K-u | nikhil-k-u.github.io

Summary

Software engineer who builds LLM applications and agents (LangChain, RAG, OpenAI and Gemini APIs, Claude Code) on top of 2+ years of production Java and Python backend work at KloudGin (full time since Aug 2025, after a year-long internship). Built a Gemini-powered page-understanding service for a browser agent with multi-model fallback and prompt-injection screening, and co-developed a SegFormer medical-image segmentation model (Dice 0.895).

Technical Skills

GenAI & LLMs: OpenAI API, Gemini API, LangChain, RAG Pipelines, Vector Search (Qdrant), Prompt Engineering, AI Agents, Claude Code (agentic coding)
ML & Vision: PyTorch, Hugging Face Transformers, SegFormer, OpenCV, Albumentations, scikit-learn, Pandas, Azure Computer Vision
Backend: Python, FastAPI, Flask, Java, Spring Boot, Microservices, REST APIs, SQL, Concurrency, Caching
Cloud & Tools: AWS (EC2, S3, SNS, SES), Azure, MySQL, MongoDB, Git, Linux, CI/CD, Agile/Scrum

AI Projects

LLM Browser-Agent Toolkit (Personal) | *Python, Gemini API, Claude Code, JavaScript* 2026
- Built a **Gemini-powered page-digest service** (Python HTTP server + injected JS collector) that turns live web pages into structured JSON actions for an AI agent, **18–55x smaller** than the raw page, orchestrated with **Claude Code**.
- **Cost, latency & safety:** multi-model fallback across Gemini models with per-key quota tracking and a Claude fallback, per-host result caching, JSON-only outputs, and page text treated as untrusted input, with instruction-like text flagged as possible prompt injection.

Colorectal Polyp Segmentation (Final-Year Project, team of 4) | *PyTorch, Hugging Face* 2025
- Co-developed a hybrid CNN + **SegFormer (MiT-b0)** Transformer model fusing local boundary features with global context; trained on Kvasir-SEG with augmentation and a Dice + BCE loss to a **Dice score of 0.895** (IoU 0.810), ahead of U-Net and TransUNet.

JalDristi – Multi-Modal AI Analytics (Smart India Hackathon, team of 4) | *Python, Azure CV, OpenAI*
- Proposed the idea and built the AI module: a Selenium social-media scraping pipeline, **Azure Computer Vision** image captions fed into **OpenAI** for crisis summaries, and a Pandas criticality score that prioritizes alerts.

Experience

KloudGin Inc. – Enterprise Field-Service & Asset Management SaaS Hyderabad, India
Associate Software Engineer *Mar 2026 – Present*
- **Data pipeline at scale:** Engineered a multithreaded Java engine that ingests hundreds of thousands of Esri GIS asset records through the KloudGin REST API, raising throughput **70%** with race conditions and duplicate processing eliminated.
- **Caching:** Designed a custom **LRU cache** that removed **300,000+ redundant API calls** per cycle, cutting batch runtime from **10+ hours to 10–20 minutes**; indexed slow queries to keep APIs under one second.

Trainee Software Engineer *Aug 2025 – Feb 2026*
- **Service portal, owned end to end:** Owned, built and deployed on **AWS EC2** a citizen-facing 311 service portal in 3–4 months, integrated with a legacy ERP through a secure server-side Backend-for-Frontend (OAuth tokens and API keys never reach the browser).
- **Cloud pipelines:** S3 image uploads with presigned URLs, and real-time SMS/email alerts on **Amazon SNS and SES**.

Software Engineer Intern – backend & integrations, converted to full time *Aug 2024 – Jul 2025*

Certifications & Training

Google Cybersecurity Professional Certificate
Microsoft Certified: AI Fundamentals (AI-900); Security, Compliance & Identity Fundamentals (SC-900)
PwC Acceleration Centers India – Cloud & Digital Launchpad (6-month training) | *Java, DevOps, Cloud* 2024

Education

B.Tech in Computer Science Engineering, CGPA 8.5 – Siksha 'O' Anusandhan University 2021 – 2025