

Vinay Kumar Agarwal

(+91) 7054118096 | vka0797@gmail.com | github.com/vinayaga97 | linkedin.com/in/vinay-kumar-agarwal

EXPERIENCE

Engineering Lead, Plotline

Bangalore, India | Sep 2025 – Present

- Leading 8 engineers at a no-code in-app engagement platform handling **1B+ requests/day**; p95 latency under 100ms
- Built 4 production agentic AI systems in Python: **support agent** (tool use, memory, structured outputs; 90% first-response accuracy), **Slack debugging agent** (auto-triage + root cause analysis + draft PRs), **query agent**, and **infra alert agent**; each with eval pipelines measuring accuracy, coverage, and fallback rates
- Designed a **RAG pipeline for event-based user segmentation**: context-aware retrieval over behavioral event schemas with LLM-generated campaign targeting; part of a broader push to make the product platform AI-native
- Shipped **AI campaign creation** using prompt chaining and structured output validation
- Built a production **MCP server** exposing Plotline's platform to customer engineering teams: read access across campaigns, segments, and analytics; write access for campaign and segment creation via tool-callable APIs
- Scaled to 9,400+ campaigns across 71 enterprise apps; expanded Off-App channels (Push, Email, SMS, RCS, WhatsApp) in 3 months; reduced AWS costs 25% and introduced org-wide observability (OpenTelemetry, Prometheus, Grafana)

Senior Software Engineer, Allen Digital

Bangalore, India | Sep 2024 – Sep 2025

- Architected an **org-wide hybrid search platform** (Go + Python) on OpenSearch: combined KNN vector search (BGE-M3 embeddings, 8k token context) with BM25 lexical search; NER-based query enrichment extracted syllabus metadata (topic, subtopic, subject) for contextual query expansion; LLM reranking layer (OpenAI, Gemini) for result verification; Kafka-driven indexing pipeline across 4 content types (Q&A, question bank, theory, multimodal/OCR)
- Led the Doubts Charter (web + app) with a 10-member team; delivered national doubt-solving platform reducing TAT from **12+ hours to 30 minutes** at scale
- Designed a declarative integration testing framework with templated code generation; cut SLT development time by **90%**

Co-Founder & Tech Lead, Cookiee App

Gurgaon, India | Mar 2023 – Aug 2024

- Led a 5-member cross-functional team to launch two consumer apps from zero; grew to 3,000+ organic users
- Built real-time backend in Go with WebSockets, MongoDB, Redis, Kafka, and FCM; achieved 30% weekly user retention
- Deployed on GCP (GKE, Cloud Run, Cloud CDN) and built a **GenAI-powered search engine** for contextual user discovery

Senior Software Engineer, Inshorts Group (PublicApp)

Noida, India | Sep 2020 – Dec 2023

- Scaled backend systems for a hyperlocal video news platform serving **10M+ DAU and 50M+ MAU**
- Led GenAI initiatives in Python: live video summaries, NER-based content tagging, duplicate detection via video frame embeddings, and automated creator behavior flagging
- Built in-house HLS conversion and streaming pipeline (FFmpeg); shipped WebRTC live streaming and native ads pipeline
- Owned 30+ services on GKE; reduced annual cloud spend by **35%, saving \$1M+**

Software Engineer, Oracle

Hyderabad, India | Jul 2019 – Aug 2020

- Built and optimized a supply chain planning engine in Java; improved production query load times by **350%**

SKILLS

Languages: Python (AI/data pipelines, FastAPI, LangChain), Go (25k+ lines, primary backend), Java, C/C++, SQL

AI/ML: Agentic AI, RAG, Context Engineering, Prompt Engineering, LangChain, LlamaIndex, LangSmith, Vector Search

Infrastructure: Distributed Systems, Microservices, AWS, GCP, Docker, Kubernetes, Kafka, OpenSearch, MongoDB, PostgreSQL, Redis, BigQuery, Terraform, CI/CD, Prometheus, Grafana, Datadog, OpenTelemetry

EDUCATION

B.Tech, Mechanical Engineering & Computer Science

Indian Institute of Technology Kanpur

2015 – 2019