

ABHISHEK KUMAR

AI Systems Engineer • Backend Engineer

abhi98166@gmail.com • +91 8091736331 • Chandigarh, India

Software engineer with 3+ years of experience designing and shipping production-grade AI systems. Specialized in agentic AI architectures, multi-agent orchestration, episodic memory, and real-time data pipelines. Proven track record of delivering end-to-end AI products — from RAG-based chatbots to fully autonomous copilots — integrated with live databases and enterprise ERP systems. Comfortable owning complete systems from architecture through deployment.

SKILLS

AI/ML: Agentic AI, Multi-Agent Orchestration, RAG, Episodic Memory, Vector Search, Prompt Engineering, LLM Tool Use, MCP (Model Context Protocol), LangChain, LangGraph, LangFuse, Human-in-the-Loop, Supervisor-Worker Patterns

Databases: PostgreSQL, MySQL, Neo4j (Graph DB), Qdrant (Vector DB), BigQuery

Languages & Frameworks: Python, Django, FastAPI, Node.js, PHP, Java, SQL

Infrastructure & Tools: Kafka, Temporal, Docker, GCP, AWS, Git, Sentence Transformers, n8n, Rundeck

EXPERIENCE

Software Engineer • Kindlife, Gurgaon

Apr 2022 – Present

► **Agentic AI Platform — Personal Copilot & Company-Wide Copilot**

- Architected and shipped a Personal AI Copilot deployed across 3 client organizations — autonomously reads and sends emails, schedules meetings, queries live data across BigQuery, MySQL, and PostgreSQL, generates analytical charts, performs deep SEO analysis, and executes actions in ERP systems via natural language instructions.
- Designed and implemented a Skills system that records user-demonstrated workflows as reusable, schedulable automations — enabling non-technical users to automate repetitive tasks (e.g. converting daily emails to Jira tickets) from a single live session with no additional configuration.
- Built a Company-Wide Copilot that handles 100% of incoming quick commerce purchase orders — parses email threads and attachments, auto-generates ERP sales orders, and routes exceptions and failures to the correct team member through the Personal Copilot.
- Designed supervisor-worker multi-agent orchestration across all three systems — a supervisor agent interprets user intent and dynamically routes tasks to specialized worker agents, each with dedicated tool sets and domain context. External service integrations built using MCP (Model Context Protocol).
- Built episodic memory pipelines using batch LLM processing of user sessions, storing structured memory in Neo4j (graph DB) and Qdrant (vector DB) for personalized, long-term context-aware responses.
- Implemented real-time triage agents that monitor AI-human conversations, detect edge cases, and escalate to human resources when needed. All systems support multimodal inputs (text and image), with Kafka for event streaming and Temporal for reliable long-running task orchestration.

► **Intelligent Customer Support Bot — 1,500+ daily conversations**

- Engineered a production GenAI customer support agent handling 1,500+ conversations daily, reducing human intervention by 50% across order queries, refunds, returns, and personalized product recommendations based on live user profile data.
- Built a self-learning, domain-specific search engine with adaptive filter discovery — the agent dynamically discovers available product filters, applies them, evaluates result quality against a threshold, and iterates until recommendations are relevant.
- Automated embedding generation and vector upsert pipelines using Python, Rundeck, and Qdrant, enabling scalable semantic retrieval with up-to-date product catalog context.

► **Performance & Infrastructure**

- Identified and eliminated ~600 redundant SQL queries through application profiling with NewRelic; implemented query caching layer via ProxySQL, reducing end-to-end application response time by 50%.
- Designed and built a CRM push notification system capable of delivering targeted, image-bearing notifications to hundreds of thousands of users simultaneously, with real-time delivery status tracking via Firebase Cloud Messaging (FCM).

Full-Stack Intern • ShopClues, Gurgaon

Apr 2021 – Apr 2022

- Integrated AI-powered recommendation engine driving 17.6% of total platform revenue through personalized product and content suggestions based on customer interaction history.
- Optimized frontend performance, improving Google PageSpeed scores from below 10% to over 90% through JS/CSS minification, image compression, lazy loading, and caching policy improvements.
- Implemented WhatsApp transactional messaging pipeline delivering real-time order confirmation, shipping, and delivery notifications to customers (Node.js, AWS SQS).

EDUCATION

B.E. Computer Science • Chitkara University, Chandigarh

2022 • 9.77 CGPA