

Abhinav Dubey

Backend-focused AI Engineer with 3+ years building production-grade LLM systems, RAG pipelines, agentic workflows and scalable cloud-native backend infrastructure for Fortune 500 clients across retail, insurance and financial services, including Ahold Delhaize, State Farm and Automobile Club of Southern California (AAA).

EXPERIENCE

TEKsystems Global Services

Hyderabad, India · Sep 2022 – Present

- Senior Software Engineer – AI** Apr 2025 – Present
 - Built a **Claude-powered IT support agent** integrated with Jira, Confluence and Slack that autonomously triages and resolves 60% of L1 tickets using tool calls and multi-turn reasoning. Median resolution dropped from 2.5 hours to ~45 minutes, saving ~200 engineer-hours/month.
 - Architected a production **RAG pipeline** (Claude + Pinecone) over 50k+ enterprise knowledge docs using hybrid search and cross-encoder re-ranking. Lifted precision@5 from 0.61 to 0.79 and reduced hallucination rate from 18% to under 4%.
 - Extended the support agent with **multi-modal capabilities** so users can paste error screenshots and stack traces. A vision-language model classifies the image, extracts text and routes the ticket; misclassification dropped ~25%.
 - Integrated **MCP servers** across 6+ operational tools (ticketing, log search, deployment monitoring, external docs APIs). Agents now autonomously pull logs, search past incidents and fetch deployment history in a single pass, replacing manual multi-dashboard triage and cutting mean time-to-root-cause significantly.
 - Leading an **SLM research initiative** with LoRA fine-tuning on 7B-class models trained on domain-specific support data. Early benchmarks show near-production accuracy at 40–50% lower inference cost and latency than larger hosted models.
- Software Engineer – AI** Apr 2024 – Mar 2025
 - Built end-to-end **NLP pipelines** using Hugging Face Transformers and spaCy for entity extraction, document summarisation and semantic search. Pipeline processes 2M+ documents per month and powers search across 6 enterprise teams.
 - Set up **LLMOps and observability** for the team's models, tracking p50/p95 latency, accuracy drift and hallucination rates in Grafana with SLO alerts. Caught 3 silent regressions before production.
 - Designed and implemented **high-throughput APIs** handling 100k+ requests per day with low latency and reliability across production workloads.
 - Implemented **gRPC-based inter-service communication** across core backend services, reducing payload size and improving inter-service response speed by 50%.
 - Deployed services on **Azure with Terraform and Kubernetes**, managing autoscaling and cluster health to maintain 99.99% uptime in production.
- Associate Software Engineer – AI** Sep 2022 – Mar 2024
 - Implemented a comprehensive **observability stack** with Prometheus, Grafana and ELK, reducing mean debugging time by 60% and giving the team full visibility into service health.
 - Automated **deployments and rollbacks** using GitHub Actions and ArgoCD, eliminating manual release steps and improving deployment efficiency across environments.
 - Implemented **circuit breaker and retry mechanisms** across core APIs, improving reliability and resilience under high load.

SKILLS

LANGUAGES & BACKEND

Python GoLang PyTorch FastAPI REST

GraphQL SQL Redis

DISTRIBUTED SYSTEMS

Docker Kubernetes Terraform gRPC

CI/CD GitHub Actions ArgoCD GitHub

CLOUD

AWS Azure Azure DevOps AWS Lambda

Bedrock

AI & LLMs

Claude API Claude Code OpenAI API

LangGraph LangChain MCP Tool Use

RAG & SEARCH

Pinecone Weaviate BM25 Hybrid Search

Re-ranking

EVALS & OBSERVABILITY

RAGAS Prometheus Grafana ELK

Langfuse

CERTIFICATIONS

Claude Certified Architect – Foundations

Anthropic

Verify certificate →

Generative AI Leader

Google Cloud

Verify certificate →

Rising Star Award

TEKsystems Global Award 2023

View award →

EDUCATION

B.Tech. Computer Science & Engineering

CSVТУ, Bhilai · Jun 2022

CGPA 9.09 / 10