

Purnendu Das

Senior Generative AI Engineer | GenAI Solution Architect | AI Platform Engineering

+91 70769 77007 | purnendu.das.in@gmail.com | LinkedIn | GitHub | Portfolio | Kolkata, India

Professional Summary

Senior Generative AI Engineer with GenAI solution architecture and AI platform leadership experience, backed by 6+ years across software engineering, data platforms, machine learning, and production AI. Designs **secure, scalable, observable, and cost-aware GenAI systems** spanning hybrid RAG, agentic and multi-agent workflows, LLM gateways, evaluation, multimodal document intelligence, and cloud-native microservices. Translates business goals into **solution architectures, non-functional requirements, API contracts, governance controls, technical roadmaps, and delivery plans**. Combines hands-on Python and cloud engineering with architecture reviews, mentoring, stakeholder communication, technical interviewing, and cross-functional delivery ownership.

Core Competencies

Architecture & Leadership	AI solution architecture, system design, NFR definition, technical strategy, architecture reviews, delivery ownership, mentoring, stakeholder management
GenAI & Agentic Systems	Hybrid RAG, agentic AI, multi-agent systems, multimodal AI, LangGraph, LangChain, LlamaIndex, prompt engineering, PEFT/LoRA
AI Platform & LLMOps	LLM gateways, model routing, evaluation, tracing, guardrails, prompt validation, PII masking, observability, auditability, cost controls, HITL
Models & Retrieval	AWS Bedrock, OpenAI, Anthropic Claude, Google Gemini, Vertex AI, OpenSearch, Pinecone, BM25, hybrid search, RRF, reranking
Backend & Distributed Systems	Python, SQL, FastAPI, Flask, Django, REST APIs, async Python, microservices, event-driven architecture, WebSockets, Pytest
Cloud & Platform Engineering	AWS, GCP, Docker, Kubernetes, Cloud Run, Lambda, Pub/Sub, AlloyDB, BigQuery, GitHub Actions, Jenkins, Airflow, Linux, CI/CD

Professional Experience

Senior Generative AI Engineer

Aug 2025 – Present

EPAM Systems

Remote, India

- Lead **solution architecture and technical design** for enterprise GenAI platforms, translating functional and non-functional requirements into service boundaries, integration patterns, API contracts, deployment models, and delivery roadmaps.
- Architect **hybrid retrieval, multimodal document intelligence, and agentic workflows** using AWS Bedrock, OpenSearch, Lambda, Python, and event-driven orchestration for secure enterprise use cases.
- Define platform controls for **scalability, reliability, security, auditability, latency, and cost**, including evaluation, guardrails, observability, retries, failure recovery, model abstraction, and controlled fallbacks.
- Provide **technical leadership** through architecture reviews, implementation guidance, coding standards, release-readiness reviews, delivery-risk resolution, and engineering mentorship.
- Partner with product, engineering, and business stakeholders; completed **50+ A3-level technical interviews** and received the **Delivery Head Award** for GenAI innovation and delivery excellence.

GenAI Associate 2

Aug 2024 – Aug 2025

PricewaterhouseCoopers Acceleration Centers (PwC AC)

Bangalore, India

- Architected **multi-agent and RAG-based microservices** on AWS, improving workflow processing efficiency by **60%** and establishing reusable AI service patterns.
- Improved retrieval accuracy by **40%** through chunking, semantic search, grounding, and response-quality optimization across enterprise knowledge workflows.
- Designed an **LLM-powered risk-intelligence workflow** with agent-based recommendations, reducing assessment turnaround time by **70%**.
- Introduced parallel execution across independent LLM chains, reducing end-to-end workflow latency by **68%** while maintaining traceability and operational controls.
- Contributed to design decisions, technical standards, release planning, production-readiness reviews, and stakeholder demonstrations.

Data Scientist

Sep 2022 – Aug 2024

Cognizant

Kolkata, India

- Led **GenAI research and productionization**, converting prototypes into secure, scalable services and reusable multi-model workflows.
- Designed solutions using **LangChain, Azure OpenAI, Gemini, Pinecone, and AWS** for semantic search, metadata enrichment, and enterprise knowledge automation.
- Built a **semantic search and RAG platform** that improved search accuracy by **35%** while reducing response latency and manual search effort.
- Established secure engineering practices using SAST, DAST, dependency scanning, and release-quality controls; achieved a **5/5 annual rating in the top 1%**.

Data Engineer

Mar 2021 – Sep 2022

Tata Consultancy Services

Kolkata, India

- Designed and implemented **scalable GCP data pipelines** using BigQuery, Dataflow, Cloud Functions, Airflow, and SQL for migration, validation, and analytics workloads.
- Reduced cloud-processing cost by **40%** and manual effort by **80%** through workflow redesign, serverless automation, and operational standardization.
- Implemented containerized CI/CD pipelines with Docker and Jenkins and won **four internal hackathons** for technical innovation.

Full Stack Web Developer (Django)

Aug 2020 – Feb 2021

Oneness Tech Solutions

Kolkata, Remote

- Developed a **predictive analytics web application** using Django and machine learning models.
- Built **REST APIs and dynamic user interfaces** connecting ML services with frontend dashboards.
- Improved website performance and SEO, increasing **organic traffic by 65%**.
- Managed the **end-to-end product lifecycle**, from database design and backend development to testing and deployment.

Selected AI & Platform Projects

Enterprise LLM Gateway | Python, FastAPI, Redis, Prometheus, Docker

GitHub

- Architected an **OpenAI-compatible, multi-tenant AI gateway** with API-key authentication, token-aware RPM/TPM limits, semantic routing, context-window protection, provider abstraction, and resilient fallback.
- Implemented **prompt validation, PII masking, tenant budgets, circuit breakers, usage and cost tracking, audit events, and Prometheus metrics**.

Enterprise Hybrid RAG Platform | FastAPI, BM25, Vector Search, RRF, RBAC/ABAC

GitHub

- Designed a **secure enterprise knowledge platform** with hybrid retrieval, reciprocal rank fusion, reranking, citation-backed answers, and access-control filtering before generation.
- Added **prompt-injection detection, PII redaction, tracing, audit events, and evaluation metrics** for retrieval quality, access leakage, refusals, citations, and latency.

EvalSeed – RAG Evaluation Dataset Generator | Python, OpenAI, CLI, JSONL

PyPi

- Built a Python library and CLI that generates **single-hop, multi-hop, and distractor Q&A pairs** from document corpora for reproducible RAG evaluation.
- Designed an **auditable filtering pipeline** using low-cost pre-filters and LLM judges for faithfulness, answerability, triviality, and difficulty.

Kitpace – Privacy-First Developer Productivity Platform | Product Architecture, Web Platform, SEO

kitpace.com

- Designed and launched a **browser-first platform with 80+ developer tools** spanning data conversion, REST/OpenAPI testing, DevOps configuration, security checks, Git workflows, and synthetic data.
- Applied a **local-first architecture** so sensitive inputs remain in the browser while supporting modular workspaces without mandatory account creation.

Publication

AI Workflow Patterns 2026 | Author & Publisher

Book Website

- Published a practical architecture guide to **production-grade AI workflows**, covering reusable patterns for RAG, agents, multi-agent orchestration, evaluation, guardrails, observability, reliability, and deployment.

Leadership & Recognition

Technical Leadership: Architecture reviews, engineering mentorship, cross-functional design alignment, technical interviewing, implementation standards, release planning, and GenAI capability building.

Recognition: EPAM Delivery Head Award; 9 PwC leadership awards; Cognizant **5/5 rating (Top 1%)**; Best Newcomer Award; 7 Cognizant leadership awards.

Professional Community: Share practical GenAI, AI architecture, and Python engineering content with a **15K+ professional audience**.

Education

B.Tech in Computer Science and Engineering

2016 – 2020

Government College of Engineering and Textile Technology, Berhampore

CGPA: 8.15/10