

PRASANT MISHRA

GenAI Solutions Architect @ IBM Labs

Bengaluru, India | (+91) 82170 50278 | mishraprasant73@gmail.com

linkedin.com/in/prasant73/ | pmishra73.github.io | github.com/pmishra73

GenAI Solutions Architect with ~11 years delivering secure, resilient, and compliant AI systems across multi-cloud and on-premise environments. Proven across the full GenAI stack — from LoRA-based fine-tuning of enterprise foundation models and synthetic data generation for LLM Training to production-grade Advanced RAG platforms, multi-agent orchestration, and MLOps at scale. A trusted technical advisor adept at designing end-to-end digital services and business solutions, translating complex enterprise requirements into scalable technical architectures. Bridges the gap between user-centric service workflows and high-scale production AI engines.

Skills

- **GENERATIVE AI & ORCHESTRATION:** LangGraph, LangChain, LlamaIndex, Multi-Agent Orchestration (Agent OS), Autonomous Loop Architectures, Semantic Kernel, Context Engineering, PEFT & LoRA-based Fine-Tuning, Synthetic Data Generation Pipelines
- **RETRIEVAL & VECTOR INFRASTRUCTURE:** Advanced RAG (Hybrid Search, Parent-Child & Semantic Chunking), Milvus, ChromaDB, Pinecone, FAISS, OpenSearch, Graph RAG, Cross-Encoder & Cohere Re-ranking, Redis Semantic Caching, Neo4j (Multi-hop Traversal)
- **ADVANCED SEARCH & LLM OPERATIONS:** RAGAS & Retrieval Evaluation Frameworks, Advanced RAG Systems, OpenSearch, Span Algebra, Multi-Stage LLM Pipelines, Production CI/CD for MLOps, Pydantic v2 Data Validation
- **ENTERPRISE MODELS & PROTOCOLS:** Claude (Anthropic), OpenAI GPT-4, Google Gemini, Mistral, MCP Server Architecture, A2A (Agent-to-Agent) Workflows, Gateway Patterns
- **AI GOVERNANCE, SAFETY AND EVALUATION:** Responsible AI Guardrails, LLM-as-Judge & RAGAS Evaluation, Bias Mitigation, Explainability/Auditability, GDPR Compliance, PII Detection, Content Safety & HAP Filters
- **SOFTWARE ENGINEERING FOUNDATIONS:** Data Structures & Algorithms, Systems Analysis, Software Architecture (Microservices, SOA, Monolithic), Design Patterns, Object-Oriented, Functional & Structured Programming Paradigms
- **PROJECT MANAGEMENT & METHODOLOGIES:** Agile, Scrum, Rapid Application Development (RAD), SDLC Governance, Architecture Review Boards (ARB), Technical Design Reviews (TDR)
- **DATA AND ML-OPS INFRASTRUCTURE:** Azure Databricks, Domino, Google BigQuery, MLflow, PySpark, Docker, Docker Compose, Kubernetes (AKS), Terraform (IaC), Apache Airflow, Spark, Kafka, Redis
- **CI/CD & DEV TOOLING:** GitHub, GitHub Actions, GitHub Workspaces, Azure Pipelines, CircleCI, Jenkins, Celery, RabbitMQ, AWS, Azure, REST, JSON, Postman
- **DATABASES:** SQL (PostgreSQL, MySQL, OracleDB, Redshift), NoSQL (MongoDB, Couchbase)
- **BACKEND & FULL-STACK:** Python (FastAPI, Django), C++, Go, Java (Spring Boot), TypeScript, Node.js, Linux Environments (RHEL, CentOS, Fedora, Ubuntu), Shell Scripting
- **CYBERSECURITY & CLOUD CONTROLS:** Basic Security Controls, Vulnerability Assessments, Data Integrity Validation, CompTIA Security+ framework, Microsoft Azure AI Fundamentals

Open Source Contributions

anchor-eval : <https://pypi.org/project/anchor-eval/0.1.0/>

Span-anchored retrieval evaluation harness for RAG pipelines over technical corpora

Python, OpenSearch, Pydantic v2, LLM APIs, Retrieval Evaluation, RAG Systems, OpenSearch, Span Algebra, LLM Pipelines, CI/CD for ML, Pydantic v2

- **Span-anchored retrieval evaluation** harness for RAG pipelines over technical corpora
Built a span-anchored retrieval evaluation framework for RAG systems over codebases, anchoring benchmark questions to raw character-level spans (doc_id, char_start, char_end) rather than chunk IDs – decoupling ground truth from chunking strategy so engineers can A/B-test chunkers, embedding models, and re-rankers without invalidating their benchmarks
- Designed a span algebra with Intersection-over-Union (IoU) scoring semantics and a 6-category archetype taxonomy (cross_file_causality, lexical_collision, negative_space, etc.) to classify retrieval failures by root cause, enabling targeted diagnosis (e.g., "chunker is too small for cross-file queries") rather than aggregate metric declines
- Implemented multi-language AST-aware chunking with per-language offset normalization (UTF-8 bytes for Go/Rust/C, UTF-16 code units for Java/Kotlin/TypeScript, Unicode codepoints for Python) across 8 languages; packaged chunker configs as signed language packs (code-oss-v0, jvm-oss-v0, systems-oss-v0) with secret-scan and tamper-detection at load time
- Architected a CI gate (anchor ci) with a 3-exit-code contract, corpus fingerprinting for drift detection, an LLM-driven question generation pipeline, BM25/dense/hybrid OpenSearch retrievers, and an embedding cache; enforced a Business Source License 1.1 gate via signed key verification – 169 unit/integration tests, PyPI release in progress

Experience

IBM – WCA4Z – GenAI Solutions Architect (Oct 2024 – Present)

- **Multi-Agent Architecture (Agent OS):** Architected and deployed a resilient multi-agent system using LangGraph and custom Model Context Protocol (MCP) servers to automate e2e enterprise code-transformation pipelines across distributed teams. Established secure A2A workflow standards, circuit breakers, and API gateways to guarantee 99.8% uptime for cross-functional agent execution.
- **LLM Fine-Tuning & Synthetic Data Generation:** Developed and deployed advanced enterprise GenAI solutions, including high-velocity Synthetic Data Generation pipelines and intelligent AI agents for workflow automation and decision support; fine-tuned IBM's Granite foundation models using LoRA-based techniques.
- **Production-Grade RAG Platform (IBM BOB):** Architected a five-stage Advanced RAG pipeline – extraction → ingestion → embedding → cross-encoder re-ranking → evaluation – to ground BOB's Explain and Transform features with domain knowledge. Built on FastAPI and LlamaIndex with semantic chunking and hybrid search over Milvus vector databases, lifting retrieval accuracy by 35% for code-domain models; deployed both on-prem and via API.
- **Document Intelligence & Extraction Engine:** Engineered a high-throughput document extraction engine (PyTorch, CNN transfer learning) classifying complex technical layouts (flowcharts vs. railroad diagrams vs. text); with OCR pipelines, n*m CSV table conversion, and multimodal captioning (Pixtral for images, Claude for tables) with Human-in-the-Loop validation feeding the RAG ingestion layer.

- **AI Governance, Safety & Observability:** Designed and enforced AI safety guardrails (profanity / HAP / non-corporate word filters), GDPR-compliant PII handling, and fully auditable outputs; formulated an offline LLM-as-a-Judge validation pipeline within the SynthGen Intelligence Board observability platform — providing real-time log analysis, tracing, and metric evaluation for an enterprise deployment scaling to 10,000+ active licenses.
- **Technical Leadership & Pre-Sales:** Provided architectural leadership and technical mentoring across IBM; led technical pre-sales and architecture reviews, pitching complex GenAI blueprints to enterprise directors and clients to align technical readiness with corporate transformation initiatives.
- **System Integrity & End-to-End QA:** Executed rigorous testing frameworks, QA protocols, and validation matrices across automated AI pipelines to guarantee absolute system/data integrity, robust edge-case handling, and enterprise-grade solution quality.

Intel – Software Application Dev Engineer & Architect (Jun 2022 – Sep 2024)

- **Graph RAG & Knowledge Systems:** Orchestrated an enterprise Knowledge Graph platform utilising Neo4j, implementing optimised multi-hop graph-traversal queries to resolve IP asset dependencies, underpinned by rigorous systems analysis of upstream data models.
- **High-Availability Ingestion:** Integrated HAProxy, Methodics, and Elastic Cloud on Kubernetes (ECK) to handle high concurrency data lookups, serving as the core structured-retrieval foundation for subsequent RAG workflows.
- **SDLC Governance & Technical Translation:** Chaired Architecture Review Boards (ARB) and Technical Design Reviews (TDR) to map intricate business requirements onto scalable technical blueprints; ensured digital service consistency and rigorous quality gate compliance across distributed, cross-functional engineering squads.

ABInBev – Data Scientist, Software Architect (Feb 2021 – Jun 2022)

- **Enterprise MLOps Platform:** Led a cross-functional team of 9 to architect an enterprise-wide MLOps data lakehouse on Azure Databricks and Domino, successfully onboarding and managing over 250+ ML models with PySpark-based feature pipelines.
- **Infrastructure Automation:** Built robust Terraform modules and Azure Pipelines to automate provisioning across Azure VNets, private endpoints, and RBAC isolation matrices, accelerating model deployment by 60%.
- **Data Lakehouse Operations:** Standardised the model lifecycle using MLflow and tracked massive analytics workloads, optimising runtime computing footprints across distributed cloud nodes.

Software Engineer & Consultant (Full-Stack / Backend Systems) (Aug 2015 – Feb 2021)

Infosys · TechMahindra · Adobe (ICS) · Brillio

- **Microservices Modernisation:** Spearheaded legacy Monolithic → Microservices transformations, decoupling core logic into RESTful Python Flask services backed by PostgreSQL/MySQL; automated deployments via Jenkins/GitHub Actions and Linux Shell/Python scripting, achieving zero unplanned downtime within an Agile/Scrum model. (Infosys)
- **Data Engineering & Workflow Systems:** Built robust data engineering solutions and lightweight enterprise workflow systems utilizing Python, C++, SQL, and event-driven architectures; managed zero-downtime microservices deployments and custom automation scripting within highly structured Linux environments.
- **Regulated Architecture & Petabyte-Scale Engineering:** Formulated fault-tolerant, SOA-aligned system blueprints using established design patterns for Tier-1 banking (Nationwide) and industrial automation (ABB) clients under strict regulatory/compliance guidelines; optimised data ingestion pipelines across ElasticDB,

Redshift, and OracleDB, resolving legacy I/O bottlenecks for telecom networks. (TechMahindra)

- **Proactive Monitoring & Reliability:** Engineered a distributed Proactive Monitoring System using RabbitMQ and Celery task queues with low-latency real-time sync, cutting production incidents by 97% across high-throughput cloud environments, backed by fault-tolerant telemetry analytics. (Adobe)
- **Real-Time Cross-Database Synchronisation (Live Sync):** Designed and built a low-latency, real-time synchronisation pipeline for Adobe Advertising Cloud, propagating data from a dual-stage backend — a pre-processing store and a downstream processed-data store — into a frontend-facing NoSQL serving layer using change-data-capture (CDC) and event-driven streaming; implemented idempotent consumers, schema validation, and conflict-resolution logic to guarantee sub-second consistency and zero data loss for live, high-throughput advertising workloads. (Adobe)
- **Edge Device Orchestration & Automation:** Programmed high-performance Resource Adapters for Network Edge Devices within distributed SD-WAN architectures, automating configurations via Blue Planet network orchestrations, Jenkins-driven CI/CD, and custom Python automation scripts to minimise manual handling errors. (Brillio)

Highest Education

Bachelor of Technology in Computer Science & Engineering (B.Tech in CSE)

West Bengal University of Technology (WBUT), 2011 - 2015 | GPA 7.21

- **Core Focus:** Systems Architecture, Core Programming, Data Structures & Algorithms.
- **Thesis Project:** Implemented digital watermarking and steganographic data hiding algorithms to protect intellectual property across distributed media files.