

Divyansh Bansal

+91 97177 98042 | divyanshb30@gmail.com | divyansh-bansal | Divyanshb30

AI Engineer building production LLM-powered agentic systems on Azure OpenAI. Deployed a **five-agent reconciliation platform** to production within Amdocs' **AT&T account**, automating **40+ hours/week**. Published in Wiley's *Software: Practice and Experience*. Databricks-certified (Data Engineering, Generative AI).

Experience

AI Software Engineer | Amdocs – AT&T

Sept 2025 – Present

New Delhi

- Architected a **five-agent reconciliation platform** (custom Python/FastAPI runtime, Azure OpenAI GPT-4.1) with per-agent cost/latency budgets, **human-in-the-loop checkpointing**, and deterministic fallbacks; all five agents **deployed and live in production**, automating **40+ hours/week** of manual work.
- Built a unified async LLM gateway with **structured-output generation**, bounded **self-repair**, **circuit-breaker failover** to deterministic logic, and per-session **cost-budget gating**.
- Designed a retrieval-grounded **NL-to-SQL pipeline** with **schema pruning**, two-layer guardrails, and dry-run **self-correction**; cut prompt tokens **~59%** and latency **~70–85%** vs. baseline.
- Built a hermetic, **CI-blocking evaluation gate** (8 offline suites, zero external dependency) validated at **1,000-reconciliation scale** across 6 failure archetypes; pinned to regression baselines.
- Implemented **four-plane agent memory** (Redis, PostgreSQL, embeddings, procedural library) driving a self-improving correction loop that promotes human-approved fixes into a reusable library.
- Leading migration of a legacy Oracle PL/SQL subscriber-event pipeline to **Azure Databricks (PySpark, bronze/silver/gold medallion architecture, Unity Catalog)**, rebuilding business-rule logic and gold-layer population to match the legacy system's outputs.
- Optimised stored procedures handling **75M records/day**, reducing latency from 45 min to 12–15 min (**~73%**). Built an **MCP-based automation agent (Apache Airflow)** enabling pause/resume/stop control over pipeline jobs during maintenance windows.
- Mentored **15+ engineers** through **Agentic AI workshops (LangChain, LangGraph)**. Presented architecture to **40+ cross-functional stakeholders**.

Co-Founder & Product Architect | DTU ERP Portal

Oct 2024 – July 2025

New Delhi

- Architected an **ML-powered ERP platform** serving **1,200+ users**; reduced accreditation-preparation time from 25 to 7 days (**~72%**) via intelligent workflows and **semantic search** (embeddings + **ChromaDB**).
- Built **8+ modules** (accreditation automation, analytics dashboards, student-alumni network); led architecture, frontend, and stakeholder adoption. Platform now **funded by DTU** for continued development.

Publication

“Design and Implementation of a Multiuser ERP Solution for Higher Education Institutions.” *Software: Practice and Experience* (Wiley)
DOI: 10.1002/spe.70060

Projects

IntelliCode: Hybrid RAG & AST Code Intelligence | HF ZeroGPU

Demo | Code

- Built a **hybrid RAG pipeline** (dense FAISS + sparse BM25, **reciprocal rank fusion, cross-encoder reranking**); Qwen2.5-3B on HF Spaces ZeroGPU. Labelled **evaluation harness** (MRR@5, Recall@k, NDCG); **121 tests** regression-gated in **GitHub Actions CI**.
- Developed async-aware **AST analyser** spanning **12 anti-pattern classes** + security scanner (injection, weak crypto, hardcoded secrets).

Loan Risk Intelligence | ~1.8M loans

API | Dash | Code

- Engineered a two-stage stacking ensemble (XGBoost + PyTorch); Test **AUC 0.9184**, validated via **McNemar's test** ($\chi^2=194.8$, $p<0.0001$), audited with **SHAP**.
- Diagnosed 2016 underwriting regime shift; deployed on **GCP Cloud Run** with live **TreeSHAP** + **Streamlit** dashboard + **PSI/KS drift monitoring**.

Decoder-Only Transformer (From Scratch) | ~10.7M params

- Built a **decoder-only transformer** from scratch with no framework abstractions: multi-head self-attention, positional encoding, layer normalisation, and autoregressive generation in raw **PyTorch**.

Technical Skills

AI/LLM: Azure OpenAI (GPT-4.1), LangChain, LangGraph, MCP, RAG, Agent Orchestration | **Retrieval:** FAISS, ChromaDB, BM25, Cross-Encoder Reranking, Semantic Search | **ML/DL:** PyTorch, TensorFlow, scikit-learn, XGBoost, LoRA/PEFT, SHAP, Transformers | **Infra:** Python, FastAPI, Docker, GCP Cloud Run, PostgreSQL, Redis, Airflow | **MLOps:** MLflow, DagsHub, CI/CD Eval Gates, Drift Monitoring (PSI/KS), Model Versioning | **Tools:** Git, GitHub Actions, Claude Code, GitHub Copilot, Pandas, NumPy, Streamlit

Education

Delhi Technological University (DTU), New Delhi

2021 – 2025

B.Tech, Electronics and Communication Engineering | **CGPA 8.06 / 10.00**

Certifications

Databricks Data Engineer Associate | Databricks Generative AI Engineer Associate