

ANURAAG MISHRA

Senior AI/ML Engineer | GenAI Systems | Retrieval & Ranking

Bengaluru, India | +91 8146928380 | anuraagmishra.214@gmail.com
linkedin.com/in/anuraagmx | github.com/mishraanuraagx | anuraagmishra.dev

Summary

Senior AI/ML Engineer with 6+ years of software engineering experience, spanning backend engineering, distributed systems, and production AI/ML systems. Currently building **RAG pipelines, retrieval, routing, semantic reuse, and evaluation layers** for regulated financial workflows, with a focus on low-latency AI services, system reliability, and safe fallback behavior. **M.Sc. Data Science** with **peer-reviewed research** combining neural networks and mathematical optimization.

Experience

Senior Associate – Technology (AI/ML)

Oct 2025 – Present

Synechron – Goldman Sachs engagement, Bengaluru

- Architected production **RAG pipelines and retrieval services** using semantic chunking, metadata schemas, query understanding, and hybrid lexical/vector retrieval, with source-scoped context construction for regulated financial workflows.
- Designed a **confidence-gated routing layer** combining deterministic rules, BM25, and semantic signals, reducing routing overhead from **2–4 s to <1 ms for ~70% of requests** while retaining safe fallback for ambiguous queries.
- Engineered low-latency retrieval paths with deterministic lookups under **20 ms** and semantic retrieval under **50 ms**, separating predictable fast paths from broader semantic retrieval when natural-language interpretation was required.
- Built **semantic reuse and query-intelligence layers** using canonicalization, entity/time constraints, hybrid retrieval, and NLI-based validation, with evaluation and safe fallback when reuse confidence was insufficient.
- Reduced embedding dimensionality from **1024D to 384D**, cutting vector storage by **62.5%**; local CPU cosine-similarity benchmarks improved by up to **2.7×**, while retrieval-quality trade-offs were evaluated before adoption.

Software Engineer

Jul 2017 – May 2022

Winshuttle (now Precisely), India

- Built production **C# backend services and REST APIs** for enterprise SAP automation, supporting distributed, high-volume data processing and long-running asynchronous workflows.
- Developed fault-tolerant **ETL pipelines processing millions of SAP records per minute**, with asynchronous execution and recovery patterns designed for large enterprise data-processing workloads.
- Reduced batch-processing time by **30%** by modernizing legacy execution components and processing flows, improving turnaround for user-triggered and long-running batch operations.
- Re-architected asynchronous processing with **RabbitMQ**, including retries, dead-letter handling, task orchestration, and restartable recovery paths with **0% observed data loss** in tracked recovery scenarios.

Selected AI Systems

Adaptive Routing – BM25, semantic signals, safe fallback, imbalance-aware evaluation.

Grounded Document Intelligence – Progressive RAG, metadata, provenance, scoped retrieval.

Semantic Cache & Reliability – Hybrid reuse, entity/time/freshness constraints, NLI, safe fallback.

Research & Publication

Research Assistant – FAU Optimization Lab

Feb 2025 – May 2025

High-quality tomographic image reconstruction integrating neural networks & mathematical optimization

Machine Learning: Science and Technology 6 (2025), 045065 | Compact ReLU edge prior → mixed-integer constraints → tomography reconstruction. CRediT – Lead: Software, Investigation, Formal Analysis, Data Curation. PyTorch, Gurobi, Slurm/HPC.

Education

M.Sc. Data Science (AI/ML), FAU Erlangen-Nürnberg, Germany – 2025

B.E. Computer Science, Panjab University, India – 2017

Technical Skills

GenAI & Retrieval: RAG Pipelines, LLMs, Agentic AI, Prompt Engineering, LangChain, OpenSearch, BM25, Hybrid Retrieval, Embeddings, Vector Databases, Semantic Caching, Query Decomposition, NER, NLI, AI Evaluation, Retrieval Evaluation.

ML & Research: PyTorch, scikit-learn, XGBoost, NumPy, pandas, Time Series, Mathematical Optimization, Gurobi, Experimentation.

Engineering: Python, C#, FastAPI, REST API Design, SQL, RabbitMQ, Backend Engineering, Microservices, Distributed Systems, System Design, AWS, Docker, Kubernetes, CI/CD, Linux.