

IBIN MATHEW BIJU

Mannheim, Germany | ibinmbiju@gmail.com | linkedin.com/in/ibin-m-biju-132245147 | github.com/ibinmbiju

PROFILE

AI engineer with 3 years of practical Python software development and 2+ years implementing and industrialising AI solutions, taking systems from requirements analysis through to production. Methodological depth in Generative AI (LLMs, RAG, agentic systems with LangGraph and LangChain), Natural Language Processing (document understanding, text classification), and MLOps (evaluation, monitoring, containerised deployment, CI/CD). Hands-on data engineering across pipelines and database systems at a scale of 10 million to 2 billion records. Works clean-code, OOP, DevOps and testing practices as standard, and has automotive industry exposure from a previous role at Mercedes-Benz. Co-author of a filed patent on multi-agent orchestration with memory and validation.

METHODOLOGICAL FOCUS AREAS

- Generative AI: LLM applications, RAG pipelines, agentic and multi-agent systems (LangGraph, LangChain, MCP), prompt engineering, evaluation and guardrails. Built and industrialised a production agent system end to end.
- Natural Language Processing: document extraction and structuring from unstructured sources, retrieval over large text corpora, hybrid search, text classification models.
- MLOps: evaluation and monitoring pipelines (Langfuse), automated validation, benchmarking and regression detection, containerised deployment with Docker and Kubernetes, CI/CD across Jenkins, GitHub Actions and GitLab.

EXPERIENCE

Agentic AI Engineering Intern

Oct 2025 – Feb 2026

NEC Laboratories Europe • Heidelberg, Germany

- Took technical responsibility for a cross-functional Generative AI initiative from requirements analysis through to production, working with domain experts, engineers and management.
- Evaluated the available agent frameworks, selected LangGraph, and designed and built the end-to-end system: orchestration, backend services, data pipelines and the interface internal users worked with.
- Industrialised the solution rather than leaving it as a prototype, including the MLOps layer: Langfuse tracing and evaluation plus an automated cross-model validator so output quality was measurable and monitored.
- Integrated tools and external data sources through MCP (Model Context Protocol), making the system extensible without rewriting agent logic.
- Enabled others by documenting architecture and decisions and building a shared knowledge hub, and presented the design and its trade-offs to upper management and the legal team for a patent filing.
- Evaluated emerging models and approaches continuously and fed what held up into the system design.

Working Student – Product (Python)

Nov 2023 – Jun 2025

AutoScout24 • Munich, Germany

- Built and managed Python data pipelines and database integrations, ingesting data from external providers into internal systems and processing 10,000+ records daily, replacing a full-day manual process.
- Implemented data validation and quality logic against messy real-world data, and operated the pipelines with CI/CD (Jenkins, Git), AWS S3 and Datadog monitoring.
- Worked directly with the Data Science team and business stakeholders on requirements and presented results.

Working Student – Software Integration

Apr 2023 – Sep 2023

Mercedes-Benz AG • Stuttgart, Germany

- Automotive software integration: built Python and Shell automation replacing manual integration steps, and carried out ECU flashing and testing on vehicle hardware.
- Collaborated with cross-functional hardware and software teams and tracked work through Git and JIRA.

Software Engineer

Jul 2021 – Nov 2022

Optum Global Solutions (UnitedHealth Group) • Hyderabad, India

- Designed and built distributed microservices and REST APIs (Java Spring Boot) processing, transforming and loading data at a scale of 10 million+ records, reaching 2 billion on one project, into relational databases.
- Applied OOP, clean code and testing practices, deployed with Docker and Kubernetes on Azure and GCP, and owned production monitoring and incident resolution.
- Coached and mentored 2 junior developers through code reviews and design discussions, and presented technical solutions to business analysts and management.

PROJECTS & RESEARCH

M.Sc. Thesis — Simulation-Based Data Generation and ML Model Development *ongoing*

- Building a physics-based simulation to generate training data, then training and iteratively optimising an ML model on it, covering the full loop from data generation to evaluation.

PubMed RAG Question-Answering System *github.com/lbinmbiju/rag_pubmed_qa*

- Built a retrieval pipeline over a large document corpus, then benchmarked embedding models, vector databases and retrieval strategies systematically. Hybrid search produced the largest gain, not a larger model.

Py4DGeo — Spatial Indexing for 3D Point Cloud Data (C++ / Python) *open-source contribution*

- Implemented a performance-critical octree structure for querying large 3D sensor datasets, reducing query runtime by 50%, validated through systematic benchmarking.

Patent (Filed, Pending) — Multi-Agent Orchestration with Memory and Validation

- Co-author on a filed patent covering orchestration of multiple AI agents with shared memory and built-in validation.

SKILLS

Languages: Python (primary), Java, SQL, C++, C#, TypeScript/JavaScript (working)

Generative AI & NLP: LangGraph, LangChain, RAG, MCP, multi-agent orchestration, prompt engineering, embeddings, Hugging Face Transformers; OpenAI, Anthropic, Gemini, Llama

ML Frameworks: PyTorch, TensorFlow, Keras, scikit-learn, Pandas, NumPy, OpenCV

MLOps & DevOps: Langfuse (evaluation, tracing, monitoring), benchmarking, Docker, Kubernetes, CI/CD (Jenkins, GitHub Actions, GitLab, Azure DevOps), Git, Linux, Datadog

Data Engineering: ETL and data pipelines, structured and unstructured data processing, relational databases (SQL, MySQL, PostgreSQL-style), vector databases (FAISS, ChromaDB, Pinecone), Neo4j, RDF/SPARQL, Kafka

Cloud: AWS, Azure, GCP

Practices: Clean code, object-oriented programming, testing, code review, technical documentation, mentoring

EDUCATION

M.Sc. Data and Computer Science *Oct 2022 – Present*

University of Heidelberg, Germany • Coursework: Generative AI using Transformers, Mining Massive Datasets, Text Analytics

B.Tech Computer Science and Engineering *Jun 2017 – Jul 2021*

Vellore Institute of Technology, India • GPA: 8.67/10

LANGUAGES

English: fluent • German: A2, actively working toward B1