Alwin Paul

Cottbus – Germany
+49 176 77354827 • alwin.paulpv@gmail.com • alwinpaul.me
alwin-paul • alwinpaul1

Objective

AI Engineer specializing in **LLM systems, RAG pipelines, and intelligent information retrieval**. Experienced in **LangChain orchestration, prompt optimization, and vector database architecture**. Passionate about building **production-grade AI systems** with measurable performance and real-world impact. Dedicated to deploying scalable language model solutions that drive business value.

Professional Experience

Perinet GmbH **Cottbus, Germany**
Working Student – AI & Systems Engineer *07/2024 – present*

- Architected end-to-end **Graph RAG pipeline**, implementing PDF parsing workflows with adaptive chunking to process product documentation.
- Engineered **prompt optimization workflows** with quantitative evaluation metrics (correctness, relevance, context precision) to reduce LLM hallucination rates.
- Built **hybrid retrieval system** combining **ChromaDB** vector search with **cross-encoder reranking models** to enhance context relevance and minimize retrieval noise.
- Implemented **semantic caching layer** using ChromaDB similarity search (0.85 threshold) to store query-response pairs, reducing redundant pipeline executions and response latency.
- Leveraged **RAGAS framework** to benchmark RAG performance across faithfulness, answer relevancy, context precision, context recall, and response latency metrics.
- Constructed **Neo4j knowledge graphs** to model document relationships and entity connections, enabling context-aware retrieval beyond vector search.

Tech Stack: Python, LangChain, ChromaDB, Neo4j, SentenceTransformers, Cross-Encoder Rerankers, RAGAS, Pandas, NumPy, Prompt Engineering, Git.

Education

Brandenburgische Technische Universität Cottbus-Senftenberg **Cottbus, Germany**
Master of Science in Artificial Intelligence *10/2022 – present*

Key Subjects:: Machine Learning, Deep Learning, Data Mining, Information Retrieval, Explainable ML, Computer Vision

Mahatma Gandhi University **Kottayam, India**
Bachelor of Computer Applications *2018 – 2021*

Projects

SmartTuner – GRPO Reinforcement Learning System:

GitHub:: `https://github.com/alwinpaul1/SmartTuner`

- Implemented **Group Relative Policy Optimization (GRPO)** from scratch for training small language models (135M–600M parameters) on logical reasoning tasks.
- Developed **reinforcement learning pipeline** with experience collection, advantage estimation, and PPO-clipped surrogate loss, improving accuracy from **46% to 60%**.
- Built **supervised fine-tuning system** using **LoRA adapters** for parameter-efficient training, generating synthetic reasoning data via OpenAI API.
- Engineered modular **CLI system** with configurable hyperparameters for reproducible experiments and real-time training visualization.

AI/ML Football Analysis System:

GitHub:: `https://github.com/alwinpaul1/AI-ML-Football-Analysis-System`

- Implemented **YOLO object detection** with **ByteTrack** for multi-player tracking across video frames, applying computer vision to sports analytics.
- Utilized **K-Means clustering** for team segmentation and **Optical Flow** for movement analysis, demonstrating versatility in algorithm selection.
- Built performance metric modules to quantify player statistics and tactical patterns, deriving actionable insights from spatial data.

Financial Analysis AI System:

GitHub:: `https://github.com/alwinpaul1/Financial_Crew`

- Designed **4-agent CrewAI system** (Query Parser, Code Writer, Code Executor, Code Reviewer) for automated financial analysis workflows.
- Integrated **FastMCP server** for **Cursor AI assistant** interoperability, creating extensible tools following industry-standard MCP protocol.
- Developed NLP pipeline using **Ollama's deepseek-r1:7b** to parse stock queries and generate production-ready Python analytics code.
- Leveraged **yfinance API** and **Matplotlib** to build real-time stock analysis and visualization capabilities.

Job Search Automation System:

GitHub:: `https://github.com/alwinpaul1/Job-Search-TG`

- Engineered automated data parsing pipeline using **Selenium** for continuous job data collection and processing.
- Implemented **NLP-based categorization** and deployed notification system with 24/7 operation and robust error handling.
- Demonstrated proficiency in **REST API integration**, workflow automation, and production deployment for scalable systems.

Achievements

Google Cloud Program: 08/2020 – 11/2020

- Completed advanced coursework on cloud AI, ML deployment pipelines, and distributed model serving.

Technical Skills

- **LLM Development:** LangChain, Prompt Engineering, RAG Design, Chunk Optimization, Context Evaluation, LLM Monitoring
- **Evaluation & Analysis:** RAGAS, Synthetic Data Generation, LLM Benchmarking, Faithfulness & Relevance Scoring
- **AI Frameworks:** PyTorch, TensorFlow, Transformers, SentenceTransformers, Hugging Face, CrewAI, Ollama
- **Vector & Graph Databases:** ChromaDB, Pinecone, Neo4j, FAISS
- **Programming:** Python (Expert), Git, Docker, Linux, RESTful APIs, Testing, Clean Code
- **Data Science:** Pandas, NumPy, Scikit-learn, Matplotlib
- **Machine Learning:** Supervised Learning, Reinforcement Learning, Model Fine-tuning, Explainable AI
- **Collaboration:** Cross-functional Development, Agile, Technical Documentation, Research Integration

Languages

English (C1), German (A2, improving)