ARAVIND PRADEEP

AI Engineer | Agentic Systems & Production RAG | Full-Stack GenAI
Cottbus, Germany | +49 176 67314504 | aravindpradeep001@gmail.com
linkedin.com/in/aravind-pradeepmadathinal | github.com/axon011 | aravindpradee.me

PROFESSIONAL SUMMARY

AI Engineer with an M.Sc. in Artificial Intelligence and hands-on production experience building agentic systems, multi-agent pipelines, and Retrieval-Augmented Generation (RAG) architectures integrated into IoT products. Ships end-to-end GenAI applications from prototype to deployment using Python, LangChain, LangGraph, FastAPI, Docker, Kubernetes, and cloud-native CI/CD. Differentiates through LLMOps maturity including Langfuse tracing, MLflow experiment tracking, and RAGAs evaluation, combined with full-stack capability in React and TypeScript.

PROFESSIONAL EXPERIENCE

Working Student - AI Integration & Agentic Systems

Perinet GmbH | Cottbus, Germany | Jun 2024 - Present

- Design and deploy RAG-based conversational agents using Python, LangChain, and Hugging Face Transformers, serving 20+ internal users daily via production REST APIs integrated into live IoT systems with sub-300ms response latency.
- Build multi-agent pipelines with LangGraph state machines and CrewAI role definitions; implement structured output parsing and automatic retry logic, reducing manual reporting effort by approximately 60% across multi-step LLM call chains.
- Develop backend services in Python and Golang connecting LLM workflows to real-time sensor data streams via MQTT; exposed as versioned REST APIs with Pydantic validation and structured error handling using FastAPI.
- Instrument all LLM calls with Langfuse for cost, latency, and quality tracing; track prompt and model iterations in MLflow, cutting average prompt-tuning cycle from days to hours through systematic A/B comparison.
- Containerize and deploy AI services with Docker and Kubernetes; automate CI/CD via GitHub Actions and GitLab for zero-touch, auditable releases to production environments.

Software Engineer Trainee

Cognizant Technology Solutions | India | Oct 2021 - Aug 2022

- Contributed to large-scale enterprise systems in agile teams; developed strong coding discipline, debugging practices, and cross-functional stakeholder communication skills.

PROJECTS

Multi-Agent Research & Report Pipeline | LangGraph, CrewAI, FastAPI, Docker, GitHub Actions

- Built a multi-agent system using LangGraph state machines with three specialized agents (Planner, Researcher, Writer) producing structured research reports; CrewAI role definitions enforce task boundaries and reduce hallucination in generated content.
- Exposed via async FastAPI REST API with Pydantic validation; containerized with Docker; CI/CD automated via GitHub Actions with zero manual steps from code push to deployment.

Production RAG System with Evaluation Harness | FastAPI, Qdrant, LangChain, RAGAs, MLflow

- Built production RAG question-answering API with hybrid retrieval (dense embeddings + BM25 sparse) over PDF and Markdown corpora using Qdrant vector database and LangChain; achieved sub-300ms p95 latency via async embedding pre-computation and query caching.
- Integrated RAGAs evaluation framework to benchmark faithfulness, answer relevance, and context recall; tracked all experiments in MLflow for systematic, reproducible improvement of retrieval quality.

Multilingual News NLP Pipeline | Whisper, XLM-RoBERTa, DistilBERT, FastAPI, PyTorch, MLflow

- Built an end-to-end German news processing pipeline: speech-to-text (Whisper ASR), cross-lingual Named Entity Recognition (XLM-RoBERTa), event classification (fine-tuned DistilBERT achieving 93.4% accuracy), translation (MarianMT), and abstractive summarization served via FastAPI with an interactive web dashboard.
- Chose cross-lingual NER over translate-then-NER approach achieving +13% F1 improvement and 8.4x faster inference; implemented smart VRAM caching to run 5 transformer models on a 4GB GPU with repeat-request latency of approximately 40ms. All experiments tracked in MLflow.

Full-Stack GenAI Study Assistant | FastAPI, React, Azure App Service, GPT-4o, SSE Streaming

- Consumer-facing AI application generating summaries, flashcards, and quiz questions via GPT-4o with Server-Sent Events streaming; containerized with Docker for Azure App Service, GitHub Actions CI/CD, rate limiting, and input sanitization for production-grade security.

SKILLS

AI & Agentic Systems: Agentic Systems, LangGraph, LangChain, CrewAI, AutoGen, Retrieval-Augmented Generation (RAG), ReAct Workflows, LLM Orchestration, Prompt Engineering, Multi-Agent Pipelines, Structured Output Parsing, Tool Use, Function Calling

LLMOps & Evaluation: Langfuse (Tracing, Evals, Prompt Management), MLflow (Experiment Tracking), RAGAs (Faithfulness, Relevance, Recall), A/B Prompt Evaluation, Model Monitoring

Machine Learning & NLP: Python, PyTorch, Hugging Face Transformers, scikit-learn, Natural Language Processing (NLP), Computer Vision, OpenAI API, Anthropic API, Embeddings, Fine-Tuning, Named Entity Recognition (NER), Text Classification, Deep Learning

Backend & APIs: FastAPI, Golang, REST APIs, Async Endpoints, Pydantic, MQTT, SQLAlchemy, Server-Sent Events (SSE), Streaming

Vector Databases & Retrieval: Qdrant, ChromaDB, Pinecone, pgvector, Hybrid Retrieval (Dense + BM25), Re-Indexing Workflows, Semantic Search

Cloud & DevOps: Docker, Kubernetes, GitHub Actions, GitLab CI/CD, Azure (App Service, Container Registry), AWS (EC2, S3), Linux, Infrastructure as Code

Frontend & Databases: React, TypeScript, JavaScript, PostgreSQL, MySQL, HTML, CSS

Languages: English (C1 Fluent), German (B1 Actively Improving), Malayalam (Native)

EDUCATION

M.Sc. Artificial Intelligence (Research Profile)

Brandenburg University of Technology | Cottbus, Germany | Oct 2022 - Est. Dec 2025

- Focus: Machine Learning, Computer Vision, Explainable ML, Data Mining, Neuromorphic Computing.
- Thesis: Content-Aware ViT Optimization on Edge Devices - pruning, quantization, and edge-hardware benchmarking for Vision Transformers using PyTorch.

B.Sc. Computer Application

BVM Holy Cross College | Kottayam, India | Jul 2018 - Mar 2021