Nabeel Javed

Munich, Germany

nabeeljaved944@gmail.com

Summary

AI engineer and backend systems specialist with experience designing production-grade LLM agents, LangChain/LangGraph workflows, RAG architectures, vector search systems, and fine-tuned domain models. Strong Python engineering foundations with real-world experience deploying scalable FastAPI services, multi-model routing, and automated LLM evaluation pipelines. Build AI systems that run in production, not prototypes with a focus on reliability, retrieval quality, and real business impact.

Experience

- **Careem** Remote
  Lead AI + Full Stack Engineer *May 2023 – Present*
  - **Agentic AI Systems (LangChain/LangGraph)**: Designed multi-agent workflows for static analysis, PR review, schema validation, and risk detection using LangChain and LangGraph-style orchestration patterns (state graphs, tool execution, retries, failure boundaries). Automated 60% of code review workload.
  - **RAG Retrieval Architecture**: Built a PGVector-based RAG system indexing logs, traces, diffs, and outages with hybrid search, embeddings, re-ranking, and structured metadata routing. Improved engineering root-cause discovery and reduced MTTR by 20%.
  - **Python AI Services**: Developed FastAPI microservices powering retrieval, model routing, embeddings pipelines, and agent inference. Implemented async workers, cache layers, and circuit-breakers improving reliability under high load.
  - **LLM Fine-Tuning & Prompt Optimization**: Performed supervised fine-tuning, prompt compression, system instruction optimization, and evaluation loops for internal classification, summarization, and debugging tasks — improving output consistency and reducing hallucination rates.
  - **Multi-Model Routing**: Integrated OpenAI (GPT-4/4o), Anthropic Claude, and Gemini with cost-aware routing, fallback strategies, rate limiting, and context-window optimization.
  - **LLM Evaluation & Monitoring**: Built automated eval pipelines measuring retrieval precision, latency, cost per request, hallucination cases, and model drift across real production traffic.
  - **Distributed Backend Engineering**: Built high-throughput WebSocket and event-driven pipelines, optimized p95 latency by 35%, and deployed serverless architectures (AWS Lambda, ECS, SNS/SQS).
- **Arbisoft**
  AI-Enabled Backend Developer *Jun 2020 – Apr 2023*
  - **AI-Assisted Legacy Modernization**: Built LLM-driven debugging and RAG tools mapping production failures to historical patterns, accelerating system modernization cycles.
  - **LLM Test Automation**: Generated contract tests, integration tests, and schema-alignment checks using LLMs improving coverage and significantly reducing regression bugs.
  - **Scalable Backend Systems**: Delivered high-availability services powering 5M monthly EdTech users with 99.9% uptime (Node/Python/Postgres/Redis).
  - **Performance Engineering**: Resolved bottlenecks using profiling, caching, query optimization, and distributed tracing eliminating 95% of slow-path issues.

AI and LLM Engineering Projects

- **LangGraph Multi-Agent Pipeline**: Built a LangGraph-based workflow with state nodes for retrieval, planning, tool selection, static analysis, and verification. Implemented error states, retries, human-in-the-loop checkpoints, and structured output validation.
- **End-to-End RAG Platform**: Engineered a production RAG stack with chunking strategies, embedding tuning, vector indexing (PGVector), hybrid BM25+embedding search, and Cohere reranking. Achieved significant precision lift over baseline RAG approaches.
- **Model Fine-Tuning + Evals**: Conducted supervised fine-tuning (OpenAI/Gemini) for domain-specific classification and summarization tasks. Built eval harnesses comparing accuracy, hallucination rate, and cost trade-offs across models.
- **LLM Static Analysis Engine**: Developed an agent that detects N+1 queries, concurrency risks, schema drift, unsafe API contracts, missing idempotency, caching inefficiencies, and backward-incompatible changes.
- **Real-Time ML Pricing Engine**: High-throughput system processing 500K+ daily requests with autoscaling, Python-based ML inference, Redis caching, and DynamoDB state storage.

Skills

- **AI/ML**: LangChain, LangGraph, RAG pipelines, embeddings, vector databases (PGVector, Pinecone), re-ranking, multi-agent orchestration, fine-tuning, eval pipelines
- **LLM Providers**: OpenAI (GPT-4/4o), Anthropic Claude, Google Gemini
- **Backend**: Python (primary), FastAPI, Flask, async systems, Node.js, PostgreSQL, Redis, Docker, microservices
- **Cloud**: AWS (Lambda, ECS, SQS/SNS), GCP (Cloud Run, Cloud SQL, Vertex AI)
- **Other**: Observability, distributed tracing, performance tuning, system design, testing strategies