

Abeel Ishrat

Germany | ishratabeel0@gmail.com | 016 25339587 | www.abeelishrat.com | linkedin.com/in/abeelishrat
github.com/AbeelIshrat1602

Education

Universität Trier, Trier, Germany April 2025 – April 2027
Master of Science in Natural Language Processing

Ghulam Ishaq Khan Institute of Engineering Sciences and Technology (GIKI), August 2019 – June 2023
Topi, Pakistan
BS in Computer Engineering

Experience

Werkstudent AI (Generative AI & NLP), MPR International GmbH – Frankfurt, Aug 2025 – Present
Germany (Hybrid)

- Shipped 5+ production Generative AI tools on **Google Cloud Platform (GCP)** using OpenAI APIs and open-source LLMs (GPT-4, LLaMA), delivering to enterprise clients with a days-to-weeks release cadence and 35% latency improvements
- Partnered with the **finance team** to automate **invoice generation** end-to-end, clearing a monthly backlog of 30+ customers that previously could not be billed and eliminating a recurring manual bottleneck in the billing workflow
- Engineered **end-to-end automation pipelines** with **n8n**, internal dashboards in **Retool**, and **Supabase** (Postgres/auth) backends on GCP, cutting manual operational overhead by 40% across 3+ client workflows and freeing an estimated 15+ hours/week for the team
- Built **document understanding** and **intelligent search** services over 10K+ enterprise documents using **Retrieval-Augmented Generation (RAG)**, transformer embeddings, and vector databases (ChromaDB, Pinecone), improving knowledge-retrieval accuracy by 25%
- Fine-tuned and deployed **transformer-based NLP models** (BERT, Hugging Face) for **entity extraction and classification** across 3+ domain-specific use cases, raising response accuracy by 20%
- Built production **AI agents** with **LangChain** and **LangGraph**, orchestrating 6+ specialized agents via stateful graphs and conditional routing with resilient tool use, state management, and error handling, cutting multi-step task completion time by 45%
- Developed custom **Model Context Protocol (MCP) servers** exposing 12+ internal tools and APIs to LLM agents through standardized **tool calling**, reducing new-integration time from days to under 2 hours
- Created **tool-calling / function-calling** pipelines that route requests across 10+ tools with 95% routing accuracy, incorporating 4+ agent patterns (tool-use, multi-step reasoning) after benchmarking
- Implemented **structured output generation** (JSON-schema enforcement), evaluation harnesses, and fallback logic for low-confidence responses, measuring accuracy, groundedness, latency, and cost to cut malformed outputs
- Packaged and deployed models with **Docker** on **GCP** (Cloud Run, Vertex AI) and **AWS**, optimising **provisioning for performance and scalability**, and used **observability tooling** (tracing, monitoring, alerting) to drive continuous improvement, sustaining 99.9% uptime and sub-200ms API response times
- Collaborated with cross-functional teams (consultants, engineers) to translate ambiguous business requirements into 5+ robust, AI-driven technical solutions

Artificial Intelligence Engineer, Radical AI – New York, USA (Remote) February 2024 – March 2025

- Architected 5+ Generative AI applications using **OpenAI GPT-4** and **Google Gemini**, lifting user engagement by 30% through advanced chatbots and AI-driven digital workers
- Delivered 10+ end-to-end AI projects from requirements to production, including **FastAPI** backend development, Dockerization, and cloud deployment on GCP, AWS, and Digital Ocean, each released in 2–4 weeks

- Integrated custom **RAG pipelines** with Pinecone, FAISS, and ChromaDB, improving information-retrieval accuracy and decision-making by 20% across 8+ deployments
- Engineered multi-modal **document understanding** systems combining domain-specific LLMs (Llama 2/3, PMC-Llama) with vector search over specialised text corpora, reducing query response times by 30%
- Orchestrated **multi-agent systems** with **LangChain** and structured prompt templates plus tool calling across 8+ tools, increasing response accuracy by 18% on complex reasoning benchmarks
- Hardened AI systems through unit and integration testing, driving a 15% reduction in production errors and monitoring live deployments so they ran reliably in production
- Collaborated with AI researchers to integrate 3 cutting-edge AI technologies into production, adhering to MLOps best practices (CI/CD, model versioning, observability)

Artificial Intelligence Engineer, ArchiWiz – Islamabad, Pakistan

Sept 2023 – February 2024

- Reduced project takeoff time by **65%** by building an intelligent web-scraping bot wired into an LLM pipeline, automating a previously manual estimation workflow end-to-end
- Launched 2 AI-powered internal automation tools that became core to ArchiWiz's engineering workflow, used daily and accelerating the company's AI adoption roadmap

Technical Skills

Languages: Python, TypeScript, JavaScript, C/C++ , C#, SQL, HTML, R

NLP: Transformer-Based Models, BERT, Hugging Face Transformers, Named Entity Recognition, Entity Resolution, Text Classification, Document Understanding, Intelligent Search, Fine-Tuning, spaCy, Multi-Lingual NLP

AI/ML Frameworks: PyTorch, TensorFlow, Keras, scikit-learn, pandas, NumPy, LangChain, LangGraph, OpenAI API, LLaMA, GPT-4, Google Gemini

LLMs & RAG: Prompt Engineering, Retrieval-Augmented Generation, Structured Output Generation, Embedding Pipelines, Evaluation & Groundedness Checks, ChromaDB, Pinecone, FAISS

Agentic AI: Agent Orchestration, LangGraph, Model Context Protocol (MCP) Servers, Tool/Function Calling, Multi-Agent Systems

Containerization & MLOps: Docker, GCP (Vertex AI, Cloud Run), AWS, CI/CD (GitHub Actions, Jenkins), Terraform, FastAPI, Flask, Observability & Monitoring

Automation & Tools: n8n, Retool, Supabase, GoHighLevel, Webhook Deployment, Selenium, OpenCV, DeepGram, ElevenLabs, AssemblyAI

Developer Tools: Claude Code, Cursor, GitHub Copilot, Git, GitHub, Postman, VS Code, Linux, PostgreSQL, MySQL, Firebase

Certifications

- Supervised Machine Learning: Regression and Classification | DeepLearning.AI, Stanford University
- Foundations: Data, Data, Everywhere | Google
- Programming for Everybody (Getting Started with Python) | University of Michigan

Languages

English: C1 (Advanced) | **German:** A2