

TARUN GUPTA

AI Systems Architect • LLM & MLOps Consultant • EU Relocation Ready

Prior Germany: Magdeburg • Hamburg • Düsseldorf | EU Blue Card: MSc work permit (2019–2022) + EU Blue Card (2022–Jan 2025) + re-issued (Apr 2025–Apr 2026), eligible fast-track re-issue | German A2 • English C1 • immediate availability
tarungupta.medium@gmail.com • +91 78388 68072 • linkedin.com/in/tarun-gupta-in • Portfolio: tarunlnmiit.github.io

PROFESSIONAL SUMMARY

I turn scattered AI capabilities into tools people can actually run.

- **10+ years** engineering production systems; **5+ years** specialised in **AI/ML, GenAI, LLM, MLOps** — from medical-imaging research (Universitätsklinikum Magdeburg) to applied data science (Vivere GmbH Hamburg) to **3+ years** cross-industry AI/ML consulting across **5 industries** (Utilities • Manufacturing • E-mobility • Retail • B2B).
- Most recent on-site: **8 months in Germany** (May 2025 – Jan 2026) delivering AI/ML for German clients in Retail/E-commerce, Manufacturing, E-mobility.
- Production AI with measurable outcomes: **\$100K+/month** savings on LLM invoice pipeline • voice AI agents auto-resolving **~65%** of inbound calls (multi-FTE workload absorbed, **~40%** average handle-time reduction) • **5 forecasting pipelines** from 1 codebase • **7 plug-and-play RAG chatbots** • open-source AI tools in production with **100+ GitHub stars**.
- **Cloud-agnostic MLOps**: AWS (Bedrock • SageMaker • Lambda • API Gateway • Aurora • CloudWatch • EMR), Azure (OpenAI • AI Search • Document Intelligence • Blob Storage • Container Apps • DevOps), Snowflake (Cortex • Snowpark • Model Registry • Feature Store • Container Services). Certified: SnowPro Core • SnowPro Specialty Gen AI • Anthropic: Intro to Agentic Skills • Building with the Claude API • Intro to MCP • Claude Code in Action • In progress: **AWS ML Specialty** (Q3 2026).

TECHNICAL SKILLS

Languages: Python (10+ yrs) • SQL • Scala • JavaScript / Node.js

AI/ML & GenAI: LLMs • RAG • Prompt Engineering • LangChain • LangGraph • CrewAI • DSPy • TruLens • PyTorch • TensorFlow • scikit-learn • YOLO • OpenCV • XAI • NLP • Time-Series • Anomaly Detection • Active Learning • Voice AI (STT/TTS)

Cloud & MLOps: AWS (Bedrock • SageMaker • Lambda • API Gateway • Aurora • CloudWatch • EC2 • S3 • EMR • Glue • Athena • EKS • MSK • Kinesis • SQS/SNS) • Azure (OpenAI • AI Search • Document Intelligence • Blob Storage • Container Apps • DevOps) • Snowflake (Cortex • Snowpark • Model Registry • Feature Store • Container Services • Tasks • Streams) • Kubernetes • Docker • Terraform • GitHub Actions • GitLab • Jenkins

Interfaces & Data: Streamlit • FastAPI • Django • Next.js • PostgreSQL • MySQL • MongoDB • BigQuery • Pandas • Apache Spark • Apache Airflow • Kafka

PROFESSIONAL EXPERIENCE

Managing Consultant — AI & Data Science • LPDG Group

Apr 2023 – Present • Germany & India • Utilities • Manufacturing • E-mobility • Retail • B2B

Voice AI Agents for Customer Service — Retail / E-commerce • Germany

- Deployed end-to-end voice AI agents (STT → LLM → TTS) handling inbound customer service calls autonomously; integrated with order management + inventory via tool calls for full resolution without human escalation. **~65% calls auto-resolved** • multi-FTE workload absorbed • **~40% average handle-time reduction**.
- **Stack**: Python • CrewAI • VAPI • n8n • Twilio • Azure OpenAI • Azure STT/TTS • Streamlit.

Intelligent Invoice Entity Recognition Platform — Utilities • India — \$100K+/month savings

- Built FTP → Azure Document Intelligence → LLM → Snowflake → Streamlit pipeline processing multi-format PDFs (**1–1,000 pages**); **confidence-gated human review** prevented silent errors; service-address normalisation against ground-truth source.
- **Stack**: Python • Snowflake • Azure Document Intelligence • Azure OpenAI • Streamlit • GitHub Actions CI/CD.

Reusable MLOps Forecasting Framework — 5 Use Cases, 1 Codebase — Manufacturing & E-mobility • Germany

- One parameterised pipeline (pre-process → train → infer → post-process) serving **5 assembly-line + e-mobility forecasting use cases** with **zero code duplication**; Snowflake Model Registry + Feature Store; GitHub Actions CI/CD across all 5 pipelines.
- Eliminated costly **Snowflake → Databricks → Snowflake round-trip**; resolved missing sensor time-series via behavioural clustering — enabling reliable forecasting where previously impossible.

7 Production RAG Chatbots — 1 Plug-and-Play Framework — Multi-client

- Shipped **7 production chatbots** from **one client-agnostic framework** (new client = new config, not new codebase); multi-turn conversation via LangChain Retrieval Agents + Conversation Memory; FAISS + Azure AI Search for semantic retrieval; FastAPI async; Kubernetes auto-scaling + Terraform-provisioned infra. Led cross-functional team of **4 engineers + 3 client SMEs** across all 7 deployments.

Applied Data Scientist • Vivere GmbH

May 2022 – Feb 2023 • Hamburg, Germany • 9-month fixed-term contract

- Built **LLM-powered product intelligence dashboard** (Azure OpenAI + BigQuery + PostgreSQL) — non-technical product teams query market data conversationally; competitor clustering, organic-vs-inorganic spend attribution, event-driven Asana/email alerting via FastAPI/Pydantic.

Data Scientist (Research) • IBMI, Universitätsklinikum Magdeburg

Oct 2020 – Mar 2022 • Magdeburg, Germany

- Active learning on BraTS2017 medical imaging: Query-By-Committee (KLD • JSD • Entropy) achieved **7% improvement** in informative sample selection over random; precision fluctuation reduced **12% → 1%** — cutting medical-image annotation cost.
- Explainable Deep Reinforcement Learning: Dropout DQN + saliency maps on MNIST/Space Invaders; **3–6% performance improvement** using only **2.5–10% training data** via MMD-Critic selection — demonstrating data efficiency over data volume.
- **IRTEX** — Explainable Image Retrieval: heatmaps + saliency maps at set-level and individual level; **80% majority user approval (85% global, 65% local explanations)** validated via user studies.
- **MSc thesis:** Human Brain Simulation (VisMem cognitive model on Nengo) • **IEEE paper:** IRTEX — Image Retrieval with Textual Explanations (XAI), ieeexplore.ieee.org/document/9582666 .

Software Engineer — Data & Backend • 3DQR GmbH

Apr 2019 – Sep 2020 • Magdeburg, Germany

- Automated MySQL schema migration framework supporting diverse configurations; **200+ end-to-end test suite** (Mocha/Chai) covering all server-side paths — enabled confident ship cadence without manual regression.

Software Engineer — Data & Backend • Careers360 • ConveGenius Group

Jan 2016 – Oct 2018 • Delhi / Noida, India

- LMS delivery managing team of 3 (Django • Node.js • MySQL • AWS); built **24x7 real-time student-query chat system** ; secure Android REST APIs with server-side business logic; analytics dashboarding for student-engagement and platform-health metrics.

EDUCATION

MSc Digital Engineering • Otto-von-Guericke-Universität Magdeburg, Germany • Apr 2019 – Mar 2022

BTech Computer Science • The LNM Institute of Information Technology (LNMIIT), Jaipur, India • Jul 2012 – May 2016

CERTIFICATIONS

Anthropic — Introduction to Agent Skills • **Anthropic** — Building with the Claude API • **Anthropic** — Introduction to Model Context Protocol • **Anthropic** — Claude Code in Action • SnowPro Core Certification • SnowPro Specialty: Gen AI Certification • In progress: **AWS Certified Machine Learning – Specialty** (Q3 2026)

SELECTED PERSONAL PROJECTS

autopilot-jobhunt — **Agentic Job Search Automation** • github.com/tarunlnmiit/autopilot-jobhunt — **100 GitHub stars** • **29 forks** • MIT • autonomous AI agent scanning **130+** company careers pages nightly • LLM-scores every role **0–100** against full resume (not keywords) • Telegram notifications • on-demand resume + cover letter drafting • multi-LLM orchestration: 4-model OpenRouter fallback chain (**zero API cost**) • MCP server listed on **Glama** (publicly hosted for Claude Code/Desktop) • pip-installable (**PyPI**) • deployed on **Raspberry Pi** • Docker + GitHub Actions CI/CD.

MindMorph — **Multi-Agent Adaptive Learning Platform** • github.com/tarunlnmiit/MindMorph_Learning_Platform — **LangGraph + CrewAI** orchestrator routing learner queries across specialised agent pipelines • multi-model routing (**GPT • Claude • Gemini**) by task type/latency/cost • RAG (**Pinecone**) with Playwright/Firecrawl grounding • DSPy + TruLens prompt evaluation • Kafka analytics loop • deployed on **Kubernetes (EKS)** with Prometheus/Grafana/OpenTelemetry observability.

AI Job Engine — **Job Search Automation** • github.com/tarunlnmiit/ai-job-engine — **12 portal scrapers** (LinkedIn • Greenhouse • Lever • Indeed • Naukri • Wellfound • Instahyre • ArbeitNow • Hrist • Remotive • WeWorkRemotely • Hacker News) • **Gemini Flash** fit scoring • dynamic resume tailoring per JD • Playwright one-click Easy Apply • Streamlit dashboard.

AI Content Machine — **Multi-Platform Content Automation** • github.com/tarunlnmiit/ai-content-machine — **45+** Python scripts publishing **200+ pieces/month** across **7 platforms** (YouTube • LinkedIn • Medium • Instagram • Twitter • Substack • podcasts) • **603-node** knowledge graph preventing repeat coverage • **85% reduction** in manual repurposing • Claude API + FFmpeg + n8n.

ADDITIONAL

Content & Community: 50+ technical articles on Medium • YouTube channel Breath of Data Science — practitioner guides.

Languages: English C1 • German A2 • Hindi Native