

TARUN GUPTA

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

Nearly 5 years in Germany: Magdeburg · Hamburg · Düsseldorf (Mar 2019 – Mar 2023 · May 2025 – Jan 2026) | 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
tarungupta.medium@gmail.com · +91 78388 68072 · linkedin.com/in/tarun-gupta-in · github.com/tarunlnmiit · Portfolio: tarunlnmiit.github.io

PROFESSIONAL SUMMARY

- 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: 9 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 197 GitHub stars across 5 shipped tools.
- 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

04.2023 – 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

05.2022 – 02.2023 · Hamburg, Germany · 9-month fixed-term contract

- Built LLM-powered product intelligence dashboard (Azure OpenAI + BigQuery + PostgreSQL) across a 500+ product catalogue — 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

10.2020 – 03.2022 · 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

04.2019 – 09.2020 · 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

01.2016 – 10.2018 · Delhi / Noida, India

- Delivered an LMS serving 80,000 students, managing a team of 3 (Django · Node.js · MySQL · AWS); built a 24×7 real-time student-query chat system reaching 50,000 users; 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 (Open Source) · github.com/tarunlnmiit/autopilot-jobhunt — 197 GitHub stars · 49 forks · MIT · 90K+ views on Instagram ([instagram.com/p/DZhkDvott2](https://www.instagram.com/p/DZhkDvott2)) · autonomous AI agent scanning 130+ company careers pages nightly and scoring 1,000+ jobs daily · LLM-scores every role 0–100 against the full resume, not keywords · Telegram notifications · on-demand tailored resume + cover letter drafting · 4-model OpenRouter fallback chain (zero API cost for end users) · MCP server listed on the official MCP Registry, Glama (Quality A) and Smithery · pip-installable (PyPI) · deployed on Raspberry Pi · Docker + GitHub Actions CI/CD.

inbox-to-action — Agentic Inbox Triage Engine (Open Source) · github.com/tarunlnmiit/inbox-to-action — One-pass agentic pipeline classifying, summarising, extracting tasks and drafting replies across an entire inbox in a single run · provider-agnostic BYOK across 6 LLM providers (Ollama fully local · OpenAI · Anthropic · OpenRouter · NVIDIA NIM · Claude Code) · draft-only — no send scope exists in the codebase, enforced by automated test · MCP server + Claude Code skill · pip-installable (PyPI) · listed on the official MCP Registry · Docker/GHCR · GitHub Actions CI.

AI Job Engine — Compliance-First Job Discovery & Scoring Pipeline (private) — Declarative scraping engine where a job portal is a YAML entry rather than a bespoke module — three fetch modes (HTTP · headless browser · human-assisted) behind one parser path, so a single offline fixture test covers all three · 198-company ATS registry with automatic vendor detection across Greenhouse · Lever · Ashby · Workable (87 resolved, 8,010 live postings) · 74 verified job portals across 16 countries · robots.txt compliance enforced in code, not documented: forbidden params asserted against both the constructed URL and the post-redirect URL · LLM fit scoring (Claude Sonnet) against the full CV with versioned rubrics so scores stay comparable across prompt changes · per-country scoring budget so low-priority countries are not starved · SQLite store, Streamlit triage, human-in-the-loop apply that fills only unambiguous fields and provably never submits — enforced by an AST test, not convention · 630+ tests.

fable-mode — Behavioral Discipline Layer for Claude Code (Open Source) · github.com/tarunlnmiit/fable-mode — SessionStart + SubagentStart hooks injecting an always-on behavioural ruleset across Opus/Sonnet/Haiiku · five recurring LLM failure modes mapped to five enforced rules, including a stronger-model consult trigger on Haiiku/Sonnet · zero-config, on by default · distributed via the Claude Code plugin marketplace · MIT.

statusline-plus — Model-Agnostic Claude Code Statusline (Open Source) · github.com/tarunlnmiit/statusline-plus — Single jq pass reading the full status-line JSON once per render instead of once per segment · 3 switchable themes (minimal · full · power) covering context usage, cost, rate limits and git branch · merge-safe settings.json patching via jq merge, with a grep/sed fallback when jq is unavailable · MIT.

AI Content Machine — Multi-Platform Content Automation · github.com/tarunlnmiit/content-machine — 150 Python scripts publishing 200+ pieces/month across 7 platforms (YouTube · LinkedIn · Medium · Instagram · Twitter · Substack · podcasts) · per-task LLM model routing from a single routing table · HyperFrames video pipeline with word-synced captions · APScheduler daemon auto-publishing from SQLite · closed-loop UTM analytics tying published content to GitHub star deltas.

MindMorph — Multi-Agent Adaptive Learning Platform · github.com/tarunlnmiit/MindMorph_Learning_Platform — LangGraph multi-agent learning platform: 14 specialised agents (orchestrator · tutor · assessment · exercise · reviewer · consensus · synthesizer · scout · adaptation · factual) composed across 6 graph workflows · mastery-gated progression with live in-browser code grading via a sandboxed executor · RAG over uploaded PDFs using local fastembed embeddings, with opt-in pgvector persistence · multi-provider LLM routing (Groq · Anthropic) · MCP tool client · FastAPI + SQLAlchemy/Alembic on PostgreSQL · Streamlit UI · Docker Compose.

ADDITIONAL

Content & Community: 252 published articles on Medium; top technical pieces have drawn 500K+ lifetime views, led by Data Preprocessing in Python at 148K · YouTube channel Breath of Data Science — practitioner guides.

Languages: English C1 · German A2 · Hindi Native