Mustafa Kerem Kurban

Geneva, Switzerland · +41 77 927 96 06 · keremkurban@proton.me
linkedin.com/in/kerem-kurban-5a40a1117 · keremkurban.site

Forward Deployed AI Engineer | GenAI Observability & Evaluation | Agentic Systems & AI Compliance

AI/ML engineer with 8+ years building production AI systems, most recently in a forward deployed role at a large Swiss bank, leading delivery from discovery to production. Deep expertise where AI quality is measured: LLM tracing and experiment tracking (W&B, MLflow 3), calibrated evaluation pipelines (RAGAS, custom tool-calling benchmarks), and agent benchmarking against human ground-truth decisions in regulated environments. Compliance focus on ISO/IEC 42001, the EU AI Act and SOC 2 Type II, with audit-ready evaluation evidence and lineage treated as deliverables. Broad across the surrounding stack: MCP servers, hybrid graph & vector RAG, RLHF fine-tuning, and user-facing React data products. Comfortable leading technical discussions from engineers to senior executives while managing concurrent workstreams. Based in Switzerland, Swiss B permit, no sponsorship required.

WORK EXPERIENCE

Wipro (Client: Large Swiss Bank) Zurich, Switzerland
Machine Learning Engineer (Forward Deployed, Contract) *11/2025 – 06/2026*

- Embedded on-site within the client's Core Engineering Data & AI function (~60% client travel), managing concurrent engagements across twelve reconciliation domain groups plus risk and operations stakeholders; led discovery, solution framing, and iterative delivery from proof-of-concept to production. Initial six-month contract extended in recognition of contributions.
- Built a multi-agent benchmarking harness evaluating autonomous agents against historical human reconciliation decisions, integrating MLflow 3 for experiment tracking and observability, Unity Catalog for governance, and a purpose-built MCP server for governed data retrieval; established audit-ready AI quality assurance in a regulated financial environment.
- Designed and shipped an agentic ingestion pipeline (Azure Document Intelligence + vision-language models, Pydantic schema-guided extraction, parallel agentic execution with cross-validation of generated artefacts) standardizing 330 multimodal SOPs into a governed data asset, cutting manual review effort by ~90%.
- Extended the client platform with custom tooling: a React application exposing SOP coverage and automatability scores as a self-serve data product, and a domain ontology / knowledge graph linking reconciliation break types to SOPs for provenance-aware, auditable agent reasoning.
- Led a bank-wide evaluation of enterprise AI/ML platforms and model registries; proposed the solution architecture that shaped programme platform strategy; ran cross-functional enablement sessions for business and technical stakeholders, from engineers to senior executives, and mentored a junior engineer.

Neptune.ai Remote
AI Researcher (Contract) *03/2025 – 10/2025*

- Productionized LLMs (Mistral, Qwen3, Llama 3.1 8B) across on-prem and cloud environments (Ollama, NVIDIA NIM, Azure AI Foundry, AWS Bedrock) with Terraform IaC and Docker; managed experiment tracking and model versioning via MLflow.
- Built MCP servers with hybrid RAG to orchestrate interoperable multi-agent platforms (LangChain, CrewAI, SmolAgents), translating research prototypes into deployable agentic systems.
- Implemented SFT and RLHF (DPO, PPO, GRPO) with DSPy prompt optimization and PEFT (LoRA, Q-LoRA, DoRA), achieving a 210% improvement over base-model baseline on private-domain alignment tasks.

EPFL Blue Brain Project Geneva, Switzerland
Machine Learning Engineer *04/2024 – 01/2025*

- Built an LLM evaluation pipeline (RAGAS metrics + custom tool-calling benchmarks) measuring output quality and hallucination rates for scientific Q&A, defining data-quality and trust criteria and feeding results into downstream RLHF workflows.
- Designed large-scale RAG pipelines (BM25, reranking, graph retrieval) and a Neo4j knowledge graph combined with Microsoft GraphRAG, delivering an SME-facing research chatbot with measurably reduced retrieval hallucinations.
- Engineered a modular Python toolkit (Pydantic, LangChain, async APIs) for cross-modal data extraction, knowledge-graph search, and scalable LLM evaluation, reducing manual research effort by ~90%.

Scientific Software Developer & Data Scientist *03/2020 – 04/2024*

- Engineered high-performance graph algorithms for triplet-motif detection in directed weighted neural networks (~1,000x speedup); deployed Dask-parallelized network statistics enabling the world's largest hippocampus model (hippocampushub.eu), validated in PLOS Biology 2024.

Intel Istanbul, Turkey

Machine Learning Ambassador *05/2017 – 09/2018*

- Client-facing technical evangelist: deployed and demonstrated edge computer-vision models on the Intel Movidius Neural Compute Stick to technical and business audiences; co-published neuroimaging research with Stanford University (80% classification accuracy from fMRI time-series).

EDUCATION

MSc, Materials Science and Nanotechnology, Bilkent University / EPFL, Switzerland 2019 – 2021
MSc, Neuroscience, Bilkent University / EPFL, Switzerland 2018 – 2019
BSc, Molecular Biology and Genetics, Boğaziçi University, Turkey 2013 – 2017

SELECTED PROJECTS

- **Agentic SOP Intelligence Platform (large Swiss bank):** Agentic extraction platform standardizing 330 multimodal SOPs into a governed data asset: Azure Document Intelligence and vision-language models orchestrated by a local agentic harness with schema-guided structured output, parallel execution, and cross-validation of generated artefacts, feeding domain ontology and knowledge graph generation for auditable agent reasoning.
- **SOP Intelligence UI (large Swiss bank):** React/Node.js data product consuming curated silver-layer SOP data served through the platform's API endpoints and promoting it into a gold-standard product view: self-serve visibility into SOP coverage, automatability scores, and process bottlenecks for operations stakeholders.
- **Azure AI reference stack:** Production-style Azure AI architecture: Terraform IaC, FastAPI backend, Streamlit frontend, and MCP tool integration across coordinated repositories. github.com/KeremKurban/azure-ai-infrastructure
- **rag-experiments:** Systematic benchmarking of chunking, embedding, and retrieval strategies across RAG platforms to identify the best configuration per use case. github.com/KeremKurban/rag-experiments
- **neuroagent (EPFL Blue Brain, open source):** LLM agent orchestrating neuroscience tools through tool calling, connecting scientific data sources to conversational workflows. github.com/BlueBrain/neuroagent

TECHNICAL SKILLS

Languages: Python, TypeScript/JavaScript (React, Node.js), SQL, Bash, Cypher

GenAI Observability & Evaluation: MLflow 3 (tracing, evals, model registry), RAGAS, custom agent and tool-calling benchmarks, ground-truth benchmarking, hallucination measurement, LLM evaluation frameworks

AI Governance & Compliance: ISO/IEC 42001 (AI management systems), EU AI Act, SOC 2 Type II; audit-ready evaluation evidence, model and data lineage (MLflow 3, Unity Catalog), human ground-truth validation

Agentic & GenAI: MCP, LangChain, LangGraph, CrewAI, DSPy, SmolAgents, hybrid RAG (BM25, reranking, GraphRAG), FastAPI, FastMCP

Fine-Tuning & Optimization: SFT, RLHF (DPO, PPO, GRPO), PEFT (LoRA, Q-LoRA, DoRA), quantization, Hugging Face

Cloud & Infrastructure: AWS (Bedrock, SageMaker, Lambda, Fargate), Azure (AI Foundry, Document Intelligence), Docker, Terraform, GitHub/GitLab CI, Databricks (Unity Catalog)

Data & Retrieval: Neo4j, PostgreSQL, MongoDB, Elasticsearch/OpenSearch, Weaviate, Milvus, Pinecone, Redis

CERTIFICATIONS

Evaluating AI Agents with Arize Phoenix (DeepLearning.AI, 2025) · Neo4j Certified Professional (2024) · LLMOps & Fine-Tuning LLMs (DeepLearning.AI, 2024) · Federated Fine-Tuning of LLMs with Private Data (DeepLearning.AI, 2025) · AWS Bedrock (2025) · Artificial Intelligence on Microsoft Azure (2026) · Quantization Fundamentals with Hugging Face (DeepLearning.AI, 2025) · Machine Learning Professional (Databricks, in progress) · Certified Developer Associate (AWS, In progress)

LANGUAGES

English (C1) · French (B1) · Japanese (B1) · Turkish (Native)