

YASH PAREEK

AI Engineer

+41 76 244 1508 · ml.ypareek@gmail.com · [linkedin.com/in/pareek-yash](https://www.linkedin.com/in/pareek-yash) · Geneva, CH

AI Engineer with 2.5+ years designing and delivering enterprise AI agents and workflow automations in production in sensitive enterprise environments. Proven track record translating complex business requirements into scalable, compliant AI solutions — with hands-on experience in responsible AI governance, agent lifecycle management, and stakeholder-led adoption. MSc in Artificial Intelligence at USI Lugano. I hold a Permit B / Ausweis B / Permis B.



TECHNICAL SKILLS

Languages	Python · C++ · SQL · TypeScript
ML & AI	LLM APIs · RAG pipelines · MLOps · HuggingFace · LLM Finetuning · LangChain · vLLM · RAGAS · AiPerf LLM Benchmarking
Frameworks	PyTorch · JAX · HuggingFace Transformers · DeepEval · LangGraph · MLFlow
Infra	On-premise deployment · Docker · Kubernetes · GitHub Actions (CI/CD) · REST APIs
Data	Chroma · BM25 · Semantic search · Weights & Biases · SQL Server · Databricks · Grafana
Governance	EU AI Act · DLP policies · Data residency

EXPERIENCE

Intern · World Trade Organisation Geneva, CH | Mar 2026 – Present

- Designed an AI classification tool using hybrid RAG (BM25 + semantic search, RRF fusion), achieving 73.6% top-1 accuracy across 10,000+ regulatory codes at <\$0.01/query, with structured outputs and confidence-threshold routing for reliability.
- Built an agentic trade-policy surveillance system using MCP, Google ADK and LLM APIs — automating a fully manual cross-jurisdictional monitoring workflow across 7+ government sources and 6 countries and served with FastAPI.
- Benchmarked on-prem deployments of Gemma 3 and Qwen 3.5 models on T4 GPUs to document metrics (tokens/sec, TTFT, TTSL, Request latency, and Req/sec).

AI Engineer · Perimattic Inc. Bengaluru | Jan 2023 – Aug 2024

- Designed and deployed conversational agentic AI systems with Langchain pipeline handling 1M+ airport travellers integrating with legacy backend and cloud native deployment seamlessly with 99.95% uptime.
- Built enterprise AI agents on Google AI Studio grounded in Docs and SharePoint document corpora as extn., enabling non-technical business users to query internal knowledge bases through natural language — driving measurable adoption across client operations.
- Contributed to the AI Evaluation framework by integrating DeepEval to LM based solutions and regression testing in Machine Learning deployments and built automatic feedback loops by observing production logs.
- Built and deployed per-facility demand forecasting models (Facebook Prophet) for airport parking utilization across multiple facilities for a UK parking operator — enabling a dynamic pricing strategy that contributed to 15% revenue growth. Implemented automated drift monitoring (scheduled cron job) to detect degradation per model and trigger selective retraining, with MLflow for model versioning and lifecycle management.
- Supported end-user training and adoption: documented workflows, facilitated handoff sessions, and upskilled business teams to operate and iterate on deployed AI agents independently.

Junior AI Engineer · Freelance Jaipur | Mar 2022 – Jul 2023

- Architected a custom RAG solution (Qdrant vector DB) with containerised LLM inference via Docker — 95% hallucination reduction with enterprise only docs, 99.95% uptime, full GDPR compliance.
- Built a real-time computer vision system (YOLO) for automated vehicle number plate detection at an airport — reducing processing time from ~60s(manual lookup) to <5s (12x improvement).

PROJECTS

Agentic Observability & SRE MCP Platform

- MCP-based observability service orchestrating AI agents on n8n for automated incident analysis from Sentry and Prometheus logs — production-grade reliability and debuggability for agent-based systems.

LLM Fine-Tuning with QLoRA

- Fine-tuned Gemma 7B using QLoRA; deployed optimized inference engine with TensorRT-LLM and serving over vLLM.

EDUCATION

MSc, Artificial Intelligence · USI Lugano, Switzerland Sep 2024 – Present

BTech, Computer Science · Rajasthan Technical University Aug 2018 – Jul 2022