Emanuel Otel

Senior AI Engineer

Cluj, Romania | manuotel@gmail.com

manuotel.com | x.com/manuotel | linkedin.com/in/manuotel | github.com/manuotel

---

Profile

Senior AI Engineer focused on production LLM systems, RAG, AI agents, evaluation, and observability. Built and led AI products across enterprise chatbots, document parsing, job classification, voice AI, and retrieval systems. Strong Python, FastAPI, and cloud background, with a track record of turning vague AI ideas into shipped systems.

Work Experience

Senior Data Scientist - Publicis Groupe

AI Team | Cluj-Napoca, Romania | Jan 2026 - Present

Working on Spirax Sarco's MiM platform, an enterprise AI assistant for internal knowledge retrieval, support, and document-based question answering across large technical document corpora.

- Led LLM observability for the MiM platform, implementing Langfuse trace instrumentation and helping the engineering team use traces for debugging, QA, latency analysis, and root-cause investigation.
- Built an end-to-end evaluation system for MiM, including golden datasets, LLM-as-judge scoring, question-level observability reports, KPI dashboards, and benchmark exports for PowerBI/Azure SQL.
- Designed MiM vs Microsoft 365 Copilot benchmarks using isolated per-question conversations, comparing answer quality, grounding, latency, failures, and user-facing behavior across STS and WM business units.
- Generated a traceable synthetic QA dataset from ~5,800 PDFs and ~60k pages, producing ~30,000 QA pairs with lineage from business unit -> category -> document -> page -> evidence block.

- Developed corpus processing and OCR pipelines with resumable page-level registries, Azure-hosted Mistral Document AI extraction, text-quality profiling, and recovery workflows for failed or oversized PDFs.
- Optimized full-corpus ingestion from roughly one week to ~4 hours, enabling faster iteration on retrieval, evaluation, and product-quality experiments.
- Ran retrieval benchmark sweeps across chunking strategies, embedding models, FAISS dense search, BM25, hybrid RRF, MMR, and hierarchical retrieval; reported retrieval quality, context precision, context waste, and token-cost trade-offs.
- Built internal QA and benchmark inspection UIs so product owners, SMEs, and engineers could review generated questions, source evidence, MiM/Copilot answers, judge scores, and failure rationales.
- Investigated bounded agentic retrieval for MiM workflows, showing stronger source coverage and answer quality than single-retrieval baselines for multi-document sales and technical support scenarios.

Senior AI Engineer - BinarCode (Contractor)

AI Team | Cluj-Napoca, Romania | Sep 2025 - Dec 2025

AI Document Parser - Production agentic AI system transforming unstructured construction documents into validated JSON workflows.

- Designed and built Gemini-powered extraction agents for 8 entity types, converting messy PDFs and Excel files into validated structured data with Langfuse observability.
- Built a parallel AI processing pipeline handling 10k+ extraction requests with retry logic, async execution, and Pydantic-based output validation.
- Developed a FastAPI backend with async format detection, Docker-based deployment, and CI/CD workflows; built a Next.js frontend with real-time progress tracking, reasoning traces, and multi-format exports.

AI ISCO Classifier - AI-powered job classification system mapping job postings to ISCO-08 international occupation standards.

- Designed a RAG-based classification engine achieving 73.2% top-1 and 97.43% top-10 accuracy, delivering 500x cost reduction vs commercial APIs ($0.0002 vs $0.10 per classification).
- Engineered multi-language semantic search (EN, DE, FR, NL) using multiple embedding models and pgvector on PostgreSQL/Supabase across 436 ISCO occupations and 1,744 records.
- Built a production REST API with FastAPI, auth, batch processing, and cost tracking; self-hosted with Docker and Kubernetes for GDPR compliance. Reduced client costs from ~EUR2,000/month to ~EUR4/month for 20k monthly classifications.

Senior AI Engineer - RebelDot (Contractor)

AI Team | Cluj-Napoca, Romania | Feb 2025 - Sep 2025

Led AI initiatives across two major clients:

Vorwerk (Platform Team) - Supported internal AI platform centralization across multiple teams.

- Helped consolidate AI development practices across teams by defining shared tooling, observability standards, and implementation guidelines.
- Drove adoption of LLM observability via Langfuse and workflow automation via n8n across AI teams; authored internal documentation, led technical presentations, and defined integration strategy.
- Created Architecture Decision Records (ADRs) and shaped best practices for internal AI development, prompt management, and observability.
- Provided cross-team AI consultancy, unblocking teams on architecture, tooling, observability, and AI delivery practices.

E.ON - Led backend and AI development for a production-grade chatbot assisting customers with invoices, billing, and support queries.

- Led design and implementation of an LLM-based support chatbot used in production by thousands of users.
- Owned backend reliability, prompt strategy, observability, and edge-case handling for a production customer-support AI system.
- Collaborated with clients, PMs, and support teams to prioritize roadmap decisions and iterate based on production feedback.

AI Engineer - ContactLoop

R&D Team | Taipei, Taiwan | July 2023 - Feb 2025

Founding Engineer working directly with the CEO and CTO. Built and launched two AI products from scratch:

- ChatFusion - LLM-powered chatbot platform processing thousands of messages daily. Built the MVP, led deployment, and onboarded major clients, including one of the largest addiction treatment providers in the US.
- VoiceFusion - Speech-to-speech voice AI platform integrating STT, TTS, and LLMs, scaled to thousands of calls per month. Built the PoC and MVP that became the foundation for production voice AI deployments used by US lending companies.

- Built LLM agents with RAG, data extraction, call forwarding, function calling, and multi-step reasoning.
- Increased chatbot lead generation by 28%, improved engagement by 15% through A/B testing, and reduced LLM error rates by 23% via LLMOps, prompt evaluation, and curated fine-tuning workflows.
- Built internal tools for automation, data scraping, and prompt templating to accelerate development cycles.
- Worked with clients, Sales, Product, and Engineering to ship quickly, deliver measurable results, and close feedback loops.

AI Research Assistant - CCU

Wen-Nung Lie's Research Team | Chiayi, Taiwan | Aug 2022 - Jan 2023

- Implemented deep reinforcement learning models (VAE + LSTM + A3C) for visual navigation in mobile robots.
- Achieved up to 4x performance improvements through model and pipeline optimization.
- Improved image encoding for VAE-based architectures by up to 1% in data reconstruction quality.

Data Engineer - ForeFlight

Data Team | Odense, Denmark | Apr 2022 - July 2023

- Built large-scale airport data scrapers with BeautifulSoup and Playwright for 300+ airports, automating ingestion into SQL pipelines.
- Delivered internal APIs and dashboards powering aviation pricing insights and decision tools.

Automation Engineer - Emerson

Process Automation Team | Cluj-Napoca, Romania | July 2020 - July 2021

- Developed control system software for industrial automation using DeltaV.
- Designed operator interfaces and ran SATs for international clients to validate safety-critical systems.

Education

- M.Eng. Advanced Robotics Technology - University of Southern Denmark, Denmark (2021-2023)
- M.Eng. Exchange Program, Advanced Manufacturing Systems - National Chung Cheng University,

Taiwan (2022-2023)

- B.Eng. Automation Engineering - Technical University of Cluj-Napoca, Romania (2017-2021)

Technologies and Languages

AI & Machine Learning

LLMs, RAG, AI Agents, LLMOps, Evaluation, LLM-as-Judge, Prompt & Context Engineering, STT/TTS, NLP, Deep Learning, Reinforcement Learning, Data Extraction, Vector Search, FAISS, Qdrant, pgvector

Models & AI Platforms

OpenAI API, Azure OpenAI, Azure AI Foundry, Google Gemini, Microsoft Copilot, Langfuse, RAGAS

Backend Engineering

Python, FastAPI, Pydantic, Async APIs, WebSockets, PostgreSQL, Supabase, Redis, Docker, Git, Bash, CI/CD pipelines

Frontend & Tooling

TypeScript, React, Next.js, Playwright, UI Prototyping, A/B Testing, Internal Tools, Prompt Debuggers

Cloud, DevOps & Infrastructure

Azure, AWS, EC2, Cosmos DB, Azure DevOps, Linux, Docker, Kubernetes, Helm Charts, Coolify, OpenTelemetry, Logging & Tracing, Deployment Automation

Product & Collaboration

Architecture Decision Records (ADRs), Technical Documentation, Cross-team Enablement, AI Consulting, Developer Experience (DevEx), Product Strategy, Stakeholder Management

---

Projects

Distributed Job Scraping System - 500k Listings

System Design / Scraping / Postgres | Jan 2026

Built a distributed job scraping system that ingested 500k job listings with FastAPI, Python subprocess isolation, and a Postgres-as-queue architecture. Designed worker orchestration, heartbeat monitoring, duplicate elimination, live execution logs, and a web control plane for deploying and

monitoring scraper nodes. Processed 500k+ listings with zero duplicate URLs and sub-50ms indexed queries, while keeping infrastructure cost at EUR0 by self-hosting Supabase/Postgres via Coolify.

Improvements in Throat Swabbing Key-Points Detection

3D Vision / Robotics / Medical AI | Odense, Denmark | Feb 2023 - June 2023

Master's thesis in collaboration with Lifeline Robotics and the University of Southern Denmark. Enhanced 3D point-cloud perception for oral cavity swabbing with a modified PointNet pipeline. Researched AI advancements from 2D key-point regression to 3D point clouds, cut model parameters by 22.5 million, and boosted accuracy by 20% compared to the legacy system.

AI Visual Navigation Mobile Robot

Reinforcement Learning / Robotics | Chiayi, Taiwan | Aug 2022 - Jan 2023

Research project in collaboration with Wen-Nung Lie's lab at National Chung Cheng University. Improved A3C-based deep reinforcement learning for visual robot navigation by integrating spatial and channel attention into VAE image encoding and adding LSTM memory cells for better pathfinding.