



MARIJA MANEVA

AI ENGINEER

maneva.me
linkedin.com/in/manevamarija
github.com/marijamineva
marijamineva99@gmail.com

PROFESSIONAL SUMMARY

AI Engineer working as a consultant at Reply, where I build LLM-powered applications, RAG systems and agentic workflows that reach production rather than stopping at a demo. I cover the whole path, from gathering business requirements to deploying, scoring and monitoring what I ship. My focus is multi-agent orchestration, retrieval architectures and document intelligence over messy multi-format sources, on a Python, FastAPI and LangChain/LangGraph stack. Native Italian, English C1, German A2.

EDUCATION

University of Pavia, Italy

Masters Degree in Computer Engineering (Data Science) 2023 - 2025

Bachelors Degree in Bioengineering 2019 - 2022

TECHNICAL SKILLS

AI & LLM Systems

RAG, AI agents, agentic workflows, prompt engineering, semantic search, LLM evaluation & observability, LangChain, LangGraph, LlamaIndex, LiteLLM, OpenAI APIs, Langfuse, RAGAS

Backend & Data

Python, FastAPI, REST APIs, async Python, Pydantic, PostgreSQL, SQLAlchemy, Alembic, MongoDB, Neo4j, FAISS, Pinecone, MongoDB Vector Search, pytest, data pipelines, multi-format document parsing

Cloud & Dev

AWS, GCP, Docker, CI/CD, Git, GitHub, Angular, Streamlit, Gradio

EXPERIENCE

Reply / AI Engineer, Consultant Jun 2025 - Present

Delivering AI solutions for enterprise clients, including a major Italian multi-utility in the electricity and gas retail market.

Proposal Validation Agent

- Architected a multi-agent system for proposal validation, with agents for rule extraction, tagging, conflict detection and verification; review time per proposal fell from days to hours for the business analysts.
- Wrote the agents in Python with LangChain on top of a RAG layer that retrieved the applicable rules, and exposed it as a FastAPI service with OpenAI models handling extraction and conflict reasoning.
- Instrumented the flow with Langfuse tracing and scored retrieval and answer quality with RAGAS on every run, instead of reviewing output by hand.

Ticket Routing Analytics Platform

- Building a statistics platform on BigQuery that shows where the AI models behind a ticket routing system misclassify, so their accuracy can be improved from evidence rather than guesswork.

Document Intelligence: File Parser and Processor

- Engineered a hybrid extraction pipeline for PDF, DOCX and PPTX that scores page complexity and routes each page to a vision or text-only model, cutting inference costs by close to half.
- Read PDFs with PyMuPDF and Office files with python-docx, rendered harder pages through Pillow for OpenAI Vision, and processed documents concurrently with asyncio.

Skill Agent

- Built a cloud-native, multi-tenant conversational profiling system for SDLC adoption with per-tenant data isolation, now used daily by thousands of people.
- Modelled the data in PostgreSQL with SQLAlchemy, ran migrations with Alembic, served the API over FastAPI, routed GPT-4 calls through LiteLLM and built the Angular client.
- Defined the AWS infrastructure, RDS and networking included, as code with AWS CDK in Python.

Conversational and Estimation Agents

- Gave a business analyst chatbot a dual-memory design, so it converses naturally and still returns consistent estimates.
- Kept short-term context in LangChain memory and long-lived knowledge in a Neo4j graph, behind FastAPI.

Produkto / Data Science - AI Engineer Dec 2023 - Mar 2025

Logo AI Generator

- Fine-tuned a generative image model on a logo dataset, giving the team a text-to-logo tool.
- Started from pretrained Hugging Face checkpoints, trained on Google Colab using an A100 GPU and wrapped the result in a Gradio interface for a direct, user-friendly first use.