Denis Craciun

AI Engineer | +393246014086 | deniscraciun18@icloud.com | LinkedIn Profile | GitHub

PROFESSIONAL SUMMARY

AI Engineer specializing in agentic systems, Agentic RAG, and LLM fine-tuning. Builds production-grade AI solutions end to end, from QLoRA/LoRA fine-tuning and self-hosted LLM serving to multi-agent orchestration, MCP tool integration, and retrieval pipelines. Deep expertise in Python, PyTorch, and the LangChain/LangGraph and Hugging Face ecosystems, with a focus on reliable, observable, and cost-efficient deployments using open-source models. Seeking to develop scalable, high-impact AI-powered products.

EDUCATION

The Open University, *Master of Science in Finance*, Milton Keynes, United Kingdom — Oct 2025 - Oct 2027
University Of Pisa, *Bachelor of Science,* Pisa, Italy — Sep 2022 - Oct 2025

WORK EXPERIENCE

Judge.me, *AI Engineer*, London, United Kingdom — Jul 2025 – Present
Staq.io, *AI Engineer*, Dubai, United Arab Emirates — Apr 2025 – Jul 2025

- Designed, built, and deployed AI/ML models and solutions aligned with Staq's product and business requirements.
- Integrated AI capabilities into existing systems, platforms, and workflows.
- Collaborated with product, data, and engineering teams to define technical requirements and deliver end-to-end AI solutions.
- Evaluated, implemented, and optimised large language models (LLMs) and other AI frameworks.
- Monitored, maintained, and continuously improved deployed models to ensure performance, accuracy, and reliability.
- Documented technical architectures, methodologies, and processes.
- Ensured all AI solutions adhere to data privacy, security, and compliance requirements.

InnovYou, *Software & AI Engineer*, Sennori, Italy — Mar 2023 – Apr 2025

- Led AI R&D initiatives by implementing agentic RAG pipelines, QLoRA fine-tuning and MCP, improving chatbot retrieval accuracy and significantly reducing model hallucinations whilst reducing costs thanks to open-source models.

Abivet, *Software Developer Junior*, Rome, Italy — Sep 2022 – Mar 2023

- Integrated third-party RESTful APIs into mobile applications, expanding feature capabilities and decreasing hardware costs.
- Optimized application performance and memory usage, reducing crash rates and ensuring stability across legacy devices.
- Designed intuitive UI/UX workflows, improving user navigation flows and decreasing drop-off rates during onboarding.
- Streamlined troubleshooting processes, reducing bug resolution time significantly and improving overall release quality.

PROJECTS

Omni – Self-Improving Agentic Operating System

- Built a durable, observable agentic runtime that decomposes natural-language goals into task graphs, runs specialized multi-phase workflows, and verifies every result with an independent verifier before marking it done.
- Engineered a self-improving eval-gated loop that adopts changes only when the full evaluation suite stays green and auto-reverts regressions, with persistent memory, saga-based rollbacks, and pluggable OpenAI-compatible model adapters (Ollama / vLLM / llama.cpp).

DerpGPT – Conversational AI Model (From Scratch)

- Built and trained a 50M-parameter conversational language model on synthetic persona-based, multi-turn dialogue datasets using PyTorch only.
- Designed the full training pipeline and evaluated model behavior, highlighting challenges in coherence and context retention in smaller LLMs.

Sapiens – Claude Code Agent Skill for AI/ML Code

- Authored an Agent Skill that auto-triggers to explain AI/ML and LLM-engineering code in plain language, grounding every example in the user's own codebase via real file paths, functions, and line numbers.
- Designed the skill description and on-demand reference modules (codebase grounding, plain-language patterns) so Claude loads the right context only when a request matches, keeping responses focused and token-efficient.

OpenClaw – Self-Hosted AI Agent Runtime

- Built a fully containerized AI agent runtime that connects a local LLM (via Ollama) to messaging channels and equips it with tools for shell execution, web browsing, file I/O, and delegated coding through Claude Code.
- Implemented persistent key-value memory, a WhatsApp channel via Baileys, and a REST API / web UI in TypeScript, with one-command Docker Compose deployment so no data leaves the user's hardware.

AI ENGINEERING SKILLS & STACK

Languages & Frameworks: Python; PyTorch; TypeScript; LangChain / LangGraph; Hugging Face Transformers; LlamaIndex
LLMs & Agents: Agentic RAG; Agent Orchestration (ReAct, Plan-and-Execute, Supervisor / Multi-agent, Reflection & Self-critique, Routing & Prompt Chaining); MCP (Model Context Protocol); Tool / Function Calling; Agent Skills; Sandbox Execution; Persistent Agent Memory; Goal Decomposition & Task Graphs; Self-Improving Eval-Gated Loops; Prompt Engineering

Retrieval & Data: Vector DBs (Pinecone / Chroma); Embeddings & Reranking; Hybrid Search; Chunking Strategies; Synthetic Dataset Generation; ETL Pipelines
Fine-Tuning & Optimization: QLoRA / LoRA Fine-tuning; PEFT; TRL; bitsandbytes; Distributed Training (DeepSpeed ZeRO); Quantization & KV-Cache Compression; Adapter Merging; Training LLMs from Scratch; Hugging Face Hub
Inference & Deployment: Self-Hosted Serving (Ollama / vLLM / llama.cpp); OpenAI-Compatible Endpoints; GGUF Conversion & Modelfiles; Chat Templating; Inference Benchmarking; Containerized Workloads (Docker); FastAPI