



Giacomo Mantovani

Senior AI Engineer & full-stack developer

Italy · Remote · Italian (native), English (fluent)

I design and build **AI systems that run in production**: agents, RAG, document and voice AI, plus the backend and cloud around them. Four years of shipping for **banks, insurers, clinics and product companies**, as tech lead and as the client's technical contact. I join your team, work in your codebase and stay accountable until it is live.

1M+

help-desk requests a year routed by an LLM agent, 81% accuracy vs 75% human

-84%

manual work on 2,500+ financial statements a day

2M+

products searchable by conversation, every item checked against the request

-45%

AI cost per article on an automated content engine

SELECTED PROJECTS

Help-desk routing agent

Major Italian bank · tech lead, team of 4

LLM agent on Azure OpenAI, RAG and LangChain routing **1M+ requests a year** into **400+ categories**. Resolution time from 5–6 days to 2–3.

Inheritance document checks

Bank · design and backend, team of 3

Legal documents compared with the bank's own data; failed checks go to an operator with the reason. **Handling time -80%**.

AI phone receptionist

IGEA, healthcare SaaS · CTO & co-founder

Multi-tenant voice agent that books, reschedules and routes calls for clinics, GDPR-aware, with low conversational latency and cached audio to cut cost.

Financial statement extraction

Bank · pipeline architect, team of 5

OCR + LLM pipeline with ~30 validation checks per document, FastAPI and Kafka on OpenShift. **2,500+ documents a day, -84% manual work.**

Conversational product search

Estyl, fashion AI · tech lead

Hybrid vector search, structured filters and re-ranking over **2M+ products** on Google Cloud, with a strict evaluation pipeline, intent routing and per-user preferences.

AI content & finance platforms

Mokka Studios · senior AI engineer

Unattended LLM pipeline from search query to published article (**-45% cost**); AWS backend and AI financial reports for an investing platform.

EXPERIENCE

2026 – present	Senior AI Engineer · Mokka Studios — AI and backend on production products
2026 – present	CTO & Co-founder · IGEA — AI voice agent for healthcare clinics
Jan – Jul 2026	Tech Lead AI Engineer · Estyl — conversational fashion assistant
2022 – 2025	AI Engineer · Sprint Reply — generative AI for banking, insurance and finance
2015 – 2022	MSc AI & Data Engineering, BSc Computer Engineering · University of Pisa

STACK

PythonFastAPITypeScriptLangChain / LangGraph

Azure OpenAIAntropicOpenAIRAG & vector DBs

Azure Document IntelligenceVapi / LiveKitGoogle Cloud

AWSAzureCloudflareKafkaDockerCI/CD

Claude Code / Codex

HOW I WORK

- Evaluation before launch, on real cases
- Spend caps and cost per request tracked
- EU data, GDPR, audit trails, no sensitive data in logs
- Clear handover: your repo, your cloud, your brand