Alessandro Osti

Senior Backend Engineer | AI/ML Systems

Italy | Remote

alosti78@gmail.com

+39 391 708 6066

Backend & Infrastructure

Python · FastAPI · Java · MySQL · MongoDB · Docker · REST APIs

AI/ML Systems

RAG Systems · LLM Integration · Vector DBs · Semantic Search · Streaming SSE

Product & Delivery

Full Ownership · Production Systems · System Architecture · Remote Work

Profile

Senior engineer with 25 years building production software. Now focused on shipping AI-powered products.

I build end-to-end systems that actually work in production: **ARIA** (RAG-based AI assistant with multi-provider LLM orchestration, semantic search, and conversational memory), **MAOTrade** (real-time data processing platform running since 2012).

I specialize in taking AI from concept to production-ready systems. Strong in backend architecture (Python, FastAPI, MySQL), AI/ML integration (RAG, vector databases, LLM APIs),

and shipping products that scale. Looking for Senior Backend or AI Engineering roles at remote-first product companies where I can build real AI products with engineering-focused teams.

Production Systems

MAOTrade - Real-time Data Platform

2012 - Present • 13+ years in production • Full ownership

Complete platform for real-time data processing, designed, built and maintained autonomously for over a decade. Demonstrates end-to-end delivery capability and long-running mission-critical systems management.

Full-stack technical implementation:

- Backend: Python (FastAPI/Gunicorn), containerized architecture
- Database: MySQL, MongoDB for analytics and timeseries
- Frontend: Vue.js with custom charting engine (Canvas2D)
- Mobile: Native Android app for real-time monitoring
- Ops: Docker, Nagios monitoring, Google Auth integration

Demonstrated competencies:

- Real-time data processing and algorithmic decision systems
- External API integration (broker APIs, market data feeds)
- Production reliability: 13+ years continuous uptime
- Full ownership: from architecture to deployment to monitoring

Key achievement: Proven track record of designing, implementing and maintaining complex systems in complete autonomy for extended periods. Production-first mindset applicable to any domain.

Python | FastAPI | MySQL | MongoDB | Vue.js | Docker | Real-time Processing

---

ARIA - Conversational AI System

2024 - Present • Production-grade RAG chatbot

Conversational AI system for complex queries, developed autonomously as a hands-on learning project. Demonstrates AI Engineering competencies applied end-to-end from design to deployment.

Architecture and components:

- **Async backend:** Python FastAPI with async/await patterns, Motor (MongoDB async driver)
- **Multi-provider LLM:** Gemini 3 Flash primary (50% cost savings, 1M token context), Claude Sonnet/GPT-4o fallback for resilience

- **RAG implementation:** Qdrant vector database, SentenceTransformers embeddings, semantic search, hybrid retrieval
- **Streaming:** SSE (Server-Sent Events) for real-time response
- **Conversation management:** MongoDB sessions, context optimization

Implemented features:

- Multi-intent detection with structured extraction
- Block-based system for composite responses (charts, tables, text)
- Semantic caching for cost optimization
- Rate limiting with point-based quota
- Analytics pipeline (Fluentd, token usage/latency metrics)

Key achievement: Production-ready system demonstrating ability to architect complete AI solutions. Autonomous management of entire stack (LLM orchestration, vector search, streaming, monitoring) with attention to costs and scalability.

RAG Systems | Multi-LLM Architecture | FastAPI Async | Qdrant | Vector Embeddings | SSE Streaming | MongoDB

Professional Experience

Senior Backend Developer

Lynx S.p.A.

Apr 2018 - Present • Full Remote

Backend development for mission-critical applications. Experience with complex architectures requiring reliability and performance.

- Backend architectures: REST APIs, microservices, system integration
- Database optimization: Oracle/PostgreSQL query tuning
- Production support: troubleshooting, incident resolution

Java | Spring Boot | Oracle | PostgreSQL | REST APIs

Software Engineer

Various enterprise clients

1999 - 2018

19 years developing production systems for healthcare, banking and enterprise. Main projects:

- Healthcare management systems: Full-stack development on complex Oracle architectures, PL/SQL optimization, Android apps with NFC/GPS integration for attendance tracking and data collection

- Mobile development: Native Android applications with offline-first architecture, barcode scanning, geolocation, Bluetooth/NFC connectivity for IoT devices
- Enterprise integration: Co-founder of a consulting company specialized in system integration — middleware development for ERP integration (Navision), email migration tools (Lotus Notes to SharePoint/Google Workspace), multi-platform mobile solutions
- Telephony systems: Development of IVR systems and C++ drivers for telecom infrastructures, real-time applications for call management in contact centers

Transferable skills: Scalable backend architectures, production debugging under pressure, autonomous project delivery, collaboration with distributed teams, full ownership of complex systems from design to maintenance.

Technologies: Java, C++, Oracle, PL/SQL, Android, IBM enterprise stack, ERP integration.

Education

AI/ML Systems - Self-directed Learning (2024-2025)

- LLM fundamentals: transformer architectures, prompt engineering, context management
- RAG systems: vector databases, embedding models, retrieval strategies
- Production deployment: FastAPI async patterns, streaming, error handling, monitoring
- Hands-on implementation: ARIA chatbot as complete applied project (12-week structured learning)

MBA in Finance | Ateneo Borsari

Key Strengths

- **End-to-end delivery:** Proven ability to take projects from design to deployment autonomously
- **Production mindset:** Focus on reliability, monitoring, error handling, cost optimization
- **Full-stack capability:** Comfortable with entire stack when needed (backend, database, frontend, ops)
- **Fast learner:** AI transition in 12 months with production-ready project
- **Long-term reliability:** 25 years experience with mission-critical systems

Availability: Remote preferred. Hybrid acceptable (Milan/Bologna 2-3 days/week). Based in Reggio Emilia, Italy.

I authorize the processing of my personal data pursuant to Legislative Decree June 30, 2003, n. 196 and art. 13 of EU Regulation 2016/679 (GDPR) solely for the purposes of personnel research and

selection.