

ALEXIS MANSILLA

Software Developer | Software Engineer | AI/ML
alexis1614@live.com.ar • [LinkedIn](#) • [Portfolio](#) • Buenos Aires, Argentina

Summary

Software Developer (AI/ML) with over 3 years of experience in creating scalable applications. I specialize in designing **multi-agent architectures** (LangGraph, LangChain) and implementing **Retrieval-Augmented Generation (RAG)** systems. I have a proven track record of integrating LLMs (GPT-4, Gemini, Claude) and developing **Computer Vision** solutions (YOLO, MediaPipe, SAM2). I possess backend experience (Node.js, Python, Elixir) and manage **vector databases** (pgvector, MongoDB Atlas Vector Search), effectively combining product vision with in-depth technical expertise.

Technical Skills

Node.js • Express.js • Fastify • Elixir • Python • GraphQL • PostgreSQL • MySQL • MongoDB • SQL • React.js • Next.js • JavaScript • TypeScript

Experience

Grupo ROISA Buenos Aires, Argentina
Software Engineer (AI) 07/2025 - Present

- Designed a modular multi-agent architecture where multiple LLMs collaborate, reason, and execute tasks in coordination using LangGraph (agent orchestration and cognitive flows), based on LangChain (framework for chains, memory, and tools), and LangFuse for comprehensive telemetry and observability.
- Implemented hybrid RAG (Retrieval-Augmented Generation) solutions with PostgreSQL + pgvector, Chroma, and MongoDB Atlas Vector Search, enabling advanced semantic retrieval and dynamic knowledge contextualization.
- Integrated OpenAI models GPT-4/4o, Claude 3.x, Gemini (Vertex AI), and Cohere within multi-agent architectures for hybrid processing of text, voice, and multimodal reasoning.
- Developed backend services in Node.js/TypeScript deployed on AWS and Modal.com, with CI/CD using GitHub Actions, Docker containers, and serverless execution.
- Designed omnichannel communication systems using Twilio, integrating TwiML, SIP Trunks, VoIP, WhatsApp Business API, WebRTC, and WebSockets, creating intelligent voice and messaging agents for real-time support.
- Managed MCP (Model Context Protocol) Servers for secure model execution and context control, and implemented task automation with n8n, optimizing maintenance and asynchronous workflows.
- Integrated Cloudflare Workers and Vertex AI Pipelines for distributed inferences and edge processing. Created vector schemas using Prisma ORM, integrating Redis as a semantic cache.
- Applied advanced observability using LangFuse and OpenTelemetry, analyzing agent performance and decision-making. Utilized Cursor IDE, Lovable, and Windsurf to accelerate development, documentation, and AI-assisted testing.

Athletia Buenos Aires, Argentina
Software Engineer (AI/ML) 02/2025 - 07/2025

- Designed and developed a comprehensive sports performance analysis platform, applying Computer Vision (CV) and Generative AI to evaluate football training videos.
- Implemented the video processing service in Python with FastAPI, optimized for AI inference tasks.
- Developed an advanced CV pipeline using Ultralytics (YOLO) for detection and segmentation, MediaPipe for pose estimation, and SAM2 (Meta). Utilized Numpy, Torch, and CV2 for metrics extraction and computation.
- Executed CV workloads on GCP Compute Engine VMs equipped with GPUs to accelerate processing.
- Integrated the Generative API of Gemini (LLM) to interpret quantitative metrics and automatically generate qualitative reports, rankings, and athlete comparisons.
- Built the main API backend in Node.js using Fastify, employing Prisma ORM for database interaction with AWS RDS.
- Developed the client platform (frontend) using Next.js and TailwindCSS for data visualization and video uploads.
- Designed a scalable microservices architecture on AWS, utilizing Docker with AWS ECS and ECR for containerization.
- Implemented an asynchronous queue system using Redis and AWS SQS to manage video processing jobs.
- Managed cloud infrastructure, including AWS S3 for storage, DynamoDB and AWS Batch for batch processing, and CloudWatch for monitoring.

Skillfitai Buenos Aires, Argentina
Full-stack Developer 12/2024 - 02/2025

- Developed an AI-based job search platform (Stack AI).
- Implemented the backend in Python and the frontend in Next.js, with a MySQL database.
- Utilized Docker for consistent environments and TailwindCSS for responsive design.

Experience

NxLabs.io	Buenos Aires, Argentina
Software Engineer	02/2024 - 02/2025
<ul style="list-style-type: none">Designed and implemented a scalable multi-tenant SaaS architecture, ensuring data isolation for multiple clients.Built a conversational commerce experience on WhatsApp using OpenAI and Azure AI, streamlining the conversation and purchase process.Implemented a multi-tenant architecture in Elixir + Phoenix, utilizing Oban, Event Sourcing, and CQRS.Orchestrated workflows with Prompt Flow and integrated Chatwoot for omnichannel support.Developed scalable microservices and optimized deployments using Docker on AWS/GCP.	
lösen	Buenos Aires, Argentina
Full-stack Developer	09/2022 - 11/2024
<ul style="list-style-type: none">Designed and developed an automotive aftersales system using microservices and vector search.Implemented the frontend with Next.js and the backend using Node.js + TypeScript.Managed data in MongoDB and participated in sprints using the Scrum methodology.	
Udemy	Buenos Aires, Argentina
Full-stack Developer	04/2022 - 08/2022
<ul style="list-style-type: none">Contributed to the development of a shoe e-commerce platform alongside a team of four developers.Utilized the SCRUM methodology with sprint planning, daily meetings, and retrospectives.Used Trello for task management and issue tracking.Implemented unit tests with Jest and Mocha to ensure system stability.Employed Git as the primary version control and collaboration tool.Collaborated under the mentorship of a senior developer, adopting good agile development practices.	
Freelance	Buenos Aires, Argentina
Full-stack Developer	02/2020 - 04/2022
<ul style="list-style-type: none">Created custom solutions for clients, including payment integration (Stripe, PayPal) and personalized forums.Utilized Lua, Node.js, and MySQL databases with a focus on scalability.	

Projects

Neural Networks (ML)	Buenos Aires, Argentina
	03/2022 - 02/2025
Developed an interactive application with real-time voice and gesture recognition using TensorFlow and Next.js. Includes session management, custom roles, and a dashboard for healthcare professionals.	
<ul style="list-style-type: none">Technologies: Node.js, Fastify, PostgreSQL, TensorFlow.js, React, Docker, Websockets	

Education

University of Buenos Aires	03/2024 - Present
Bachelor's Degree in Computer Engineering	

Languages

English Intermediate ● ● ● ● ●