

Gonzalo Martin Perez

AI Engineer | Software Engineer | Product Engineer

LLM Agents · Agentic AI · RAG · GraphRAG · LLM Evals | LangChain · LangGraph · OpenAI API · Claude API · MCP · pgvector · Python · TypeScript · FastAPI · Next.js · AWS

gonzalomartinperez2002@gmail.com | +54 9 2984 686913 | Bahía Blanca, Argentina · 100% remoto · Inglés B2
linkedin.com/in/gonzalo-martin-perez | github.com/gonzalomartinperez | gonzalomartinperez.com

RESUMEN PROFESIONAL

Soy **AI Engineer** con más de 2,5 años llevando software a producción, últimamente con agentes LLM, sistemas RAG y suites de evaluación. Lidero la ingeniería de una startup fintech y blockchain cuya interfaz generativa acerca los activos digitales a cualquier persona, y me ocupo de su asistente de IA de punta a punta: agentes en LangGraph en tres idiomas, la suite que cada versión debe aprobar y el backend, la web y la app móvil. Antes desarrollé un asistente con GraphRAG para Payway.

EXPERIENCIA PROFESIONAL

AI Engineer | Rampy

Sep 2026 – Actualidad | Estados Unidos · Remoto

Lidero la ingeniería de una startup fintech y blockchain cuya app no custodial acerca los activos digitales a cualquier persona mediante una interfaz generativa. Colaboro con los tres fundadores y con especialistas en blockchain e integraciones.

- **Construí la suite de evaluación que cada versión de un asistente de IA multilingüe debe aprobar: 1.000+ escenarios replicados en inglés, español y portugués (3.000+ casos), con los agentes en un 90% o más.** Un LLM-as-a-judge en LangSmith, validado con revisión humana, califica cada ejecución.
- **Migré el asistente de Agno a un sistema multi-agent en LangGraph** con ciclos, herramientas financieras, procesos en segundo plano, reintentos y recuperación ante fallos; Claude es el modelo principal, con fallback a DeepInfra y Vertex AI, GraphRAG (Neo4j, pgvector) y memoria con Mem0.
- **Hice que las respuestas de los agentes fueran 2x más rápidas y reduje un 40% el consumo de tokens** al eliminar llamadas redundantes a herramientas y ajustar memoria y contexto, con verificación por pruebas de carga y evaluaciones.
- **Diseñé el contrato de salida estructurada que convierte las respuestas de los agentes en generative UI en la app en React Native:** botones, tarjetas y accesos nativos a swaps, vaults y perpetuos con datos en tiempo real. El usuario confirma cada transacción en su wallet no custodial de Privy.
- **Reestructuré el backend y el frontend heredados con Clean Architecture y reduje el tiempo de build y despliegue de alrededor de 1 hora a 10 minutos.** Sobre esa base construí las integraciones de trading y el modelo de comisiones: perpetuos de Hyperliquid (200+ mercados), swaps con LI.FI (100+ tokens) y vaults de Morpho, Aave y Compound, con seguimiento por webhooks.

Software Engineer | Teamcubation

Dic 2025 – Sep 2026 | Argentina · Remoto

Consultor en la plataforma de promociones de Payway (350.000+ comercios, 13 países).

- **Desarrollé un asistente con GraphRAG que aprobó el 95% o más de una suite de 500+ casos**, calificada en LangSmith con un LLM-as-a-judge validado por revisión humana. Usé LangGraph, Neo4j y PostgreSQL/pgvector, e incorporé guardrails contra preguntas fuera de dominio y fugas de datos.
- **Entregué el pipeline orientado a eventos que ingirió 20M+ promociones históricas en producción** (S3 y SQS hacia una Lambda en Python/FastAPI y luego servicios Java/Spring Boot y NestJS en Docker y Kubernetes). Desarrollé herramientas para reproducir cargas localmente y conciliar registros fallidos.

Software Engineer | Cooperativa Obrera

Dic 2024 – Dic 2025 | Bahía Blanca, Argentina · Presencial

Ingeniero principal del backoffice de permisos de la mayor cooperativa de consumo del país (2,7M+ asociados).

- **Integré 10+ sistemas internos en una sola experiencia para administradores** mediante un BFF en Python/FastAPI y un frontend en Next.js; una búsqueda manual de 10 minutos pasó a menos de 10 segundos.
- **Reduje en más de un 80% los tickets de permisos y las asignaciones manuales, y en más de un 95% los errores recurrentes de asignación**, al estandarizar la incorporación de nuevos sistemas.

AI Engineer & Full-Stack Developer | Independiente

Ene 2024 – Actualidad | Argentina · Remoto · Por proyecto

Proyectos de IA y full stack para clientes que venden en Shopify y Hotmart; varios ya están en producción.

- **Reduje procesos manuales de horas o días a minutos** con asistentes de WhatsApp, automatizaciones en n8n, ManyChat y Meta Ads, y agentes de contenido con MCP sobre Claude u OpenAI.
- **Entregué sitios web completos, software a medida y proyectos de e-commerce en días o semanas que generaron USD 10.000+ de ganancia para clientes**, entre ellos la tienda de Pequeverso y su asistente de compras con GraphRAG en LangGraph.

HABILIDADES TÉCNICAS

Ingeniería de IA: AI Agents, Multi-Agent Systems, RAG, GraphRAG, LLM Evaluation, LLM-as-a-Judge, Tool Calling, Generative UI, Prompt Engineering, AI Guardrails

Frameworks y plataformas de IA: LangGraph, LangChain, LangSmith (LLM Observability), OpenTelemetry, Claude API, OpenAI API, Vertex AI, MCP, Neo4j, pgvector

Backend y datos: FastAPI, NestJS, Spring Boot, Laravel, REST APIs, Microservices, Clean Architecture, Event-Driven Architecture, PostgreSQL, Redis

Lenguajes y frontend: Python, TypeScript, Java, SQL, Node.js, React, Next.js, React Native, Kotlin, Swift

Cloud y delivery: AWS Lambda, Amazon S3, Amazon SQS, Docker, Kubernetes, GitHub Actions, SSE

Fintech e integraciones DeFi: Hyperliquid, LI.FI, Morpho, Aave, Compound, Privy, Non-Custodial Wallets

PROYECTOS DESTACADOS

Portfolio Assistant · Asistente RAG bilingüe con evaluación

Sep 2026 – Actualidad

- **Asistente RAG bilingüe de código abierto que responde preguntas de reclutadores sobre mi trabajo:** FastAPI y LangGraph con recuperación híbrida léxica y vectorial más GraphRAG (pgvector, Neo4j), respuestas citadas por SSE y límites de costo por solicitud. Implementado y evaluado; todavía sin desplegar.
- **Creé un banco bilingüe de 1.000 preguntas más casos adversariales**, una evaluación de 244 turnos con el modelo real de OpenAI y 800+ pruebas en CI con GitHub Actions y trazas en OpenTelemetry.

github.com/gonzalomartinperez/portfolio-assistant-api

Filomena · Plataforma de exámenes en producción para ciencias de la salud

Ene 2025 – Dic 2025

Proyecto final de carrera que hoy usan en producción **cinco instituciones nacionales de Argentina** para exámenes de ciencias de la salud, con más de 1.000 usuarios simultáneos. Colaborador principal en un equipo de tres (Laravel, Next.js, Redis); calificación **10/10**.

gonzalomartinperez.com/es/work/filomena

EDUCACIÓN

Ingeniero en Sistemas de Información | Universidad Nacional del Sur

Ene 2020 – Oct 2025 | Bahía Blanca, Argentina | Promedio: 8,67 / 10