Gonzalo Sestopal

gonzasestopal@gmail.com | +525562021196 | Mexico City, Mexico

Summary

Senior Backend Engineer (Python/TypeScript) with 5+ years building and re-architecting high-traffic, production microservices in fintech and SaaS. Strong in distributed systems, event-driven workflows, AWS cloud infrastructure, and AI/LLM-enabled platforms. Proven ownership delivering reliable, observable systems under strict SLAs and cost constraints**.**

Experience

VisitoAI | VisitoAI   nov 2024 - Present
Founding Engineer · Full Stack Developer · Full-time   Mexico City, Mexico

- Led WhatsApp-based direct booking flows, increasing bookings **+50%** and delivering **up to 10× ROI**.
- Built and scaled **TypeScript microservices** powering conversational booking workflows, defining service boundaries and API contracts to ship quickly and safely.
- Designed AI conversational funnels (intent routing, guardrails, fallback flows) and retrieval strategies to maximize conversion quality.
- Introduced **RAG** to improve LLM accuracy, reduce hallucinations, and optimize token usage—directly improving unit economics/pricing efficiency.
- Implemented real-time updates across services via an event-driven flow (**RabbitMQ + WebSockets**) using ACK-based processing, reducing polling and improving UI responsiveness, including common DLQ and idempotency patterns.

Tech stack: Typescript, NodeJS, React, MongoDB, Azure, LLMs, RAG, LangChain

Fondeadora Empresarial | Fondeadora Empresarial   nov 2021 - nov 2024
Software Engineer · Tech Lead · Full-time   Mexico City, Mexico

- Defined the end-to-end architecture for a B2B fintech platform built on microservices, from AWS infrastructure (**ECS + Lambda**) to service boundaries and domain ownership; established CI/CD with **blue/green deployments** to avoid downtime and improve delivery velocity.
- Designed and implemented **payment orchestration** using **AWS Step Functions** and **TypeScript**, enabling resilient multi-step execution with retries, timeouts, and compensating actions.
- Implemented resilient async processing between microservices using **SQS FIFO**, with **DLQs + idempotency** to handle retries safely and prevent message loss under load.
- Built and mentored a team of 6 engineers, establishing backend standards (service templates, PR reviews, testing practices, and observability conventions) to improve consistency and reliability.

Tech Stack: Python, Typescript, PostgreSQL, Redis, AWS (ECS, Lambda, SQS), Docker, GitHub Actions

Fondeadora | Fondeadora   nov 2018 - nov 2021
Founding Engineer · Backend Engineer · Full-time   Mexico City, Mexico

- Reduced contacts read-path latency from ~500ms to ~50ms by implementing a **write-through Redis cached read model** (cache updated on contact writes), eliminating expensive query aggregation under load.
- Integrated and maintained critical external providers (card processor, KYC, and transfer rails), enabling secure onboarding, payments, and fund movements.
- Designed fault-tolerant distributed workflows using **load balancing**, durable **FIFO queues**, and **idempotent, retry-safe** consumers to ensure correctness under at-least-once delivery.

Projects

Debate Bot | [Live]   nov 2024 - Present
                                                          Generative AI + NLP

- Architected a structured debate system combining LLM generation with **NLI-based semantic verification** (entailment/contradiction/neutral) to validate claims against hypotheses.
- Integrated a **Hugging Face NLI classifier** to score argument validity and detect contradictions, improving evaluation robustness beyond self-critique.
- Designed **multi-agent orchestration** (proposer/opponent/judge) with rule-based constraints to enforce logical consistency and reduce hallucinated reasoning.
- Implemented **Redis caching**, **PostgreSQL persistence**, and **Dockerized deployment** for production-ready execution and repeatable runs.
- Added multi-provider support (**OpenAI + Anthropic**) with a unified interface for model routing, retries, and configuration.

Tech Stack: Python, Hugging Face, Redis, PostgreSQL, Docker, OpenAI/Anthropic SDKs

Restaurant Copilot | [Github]   nov 2024 - Present
                                                              Generative AI

Designed and built an LLM-powered assistant translating natural language into safe SQL analytics queries.
· Implemented Retrieval-Augmented Generation (RAG) using LangChain to improve response accuracy and relevance.
· Built evaluation pipelines and trace-based debugging using LangSmith to measure quality and cost trade-offs.
· Optimized inference pricing by reducing token usage while maintaining answer quality

Tech Stack: Python, FastAPI, PostgreSQL, LangChain, LangSmith, RAG

Skills

Skills: Python, Typescript, FastAPI, Node.js, Microservices, PostgreSQL, Redis, MongoDB, RabbitMQ, SQS, Docker, RAG, LangChain, LangSmith, LLM, AWS, Azure

Links

[LinkedIn], [GitHub]