

TONATIUJ SÁNCHEZ JIMÉNEZ

AI Full-Stack Engineer

tonatiuj.sanchez@gmail.com | +52 757 125 0112 | Idiomas: Inglés A2 con lectura técnica, Español Nativo, Nahuatl
<https://tonatiujsanchez.com> | <https://www.linkedin.com/in/tonatiujsanchez>

AI Full-Stack Engineer con más de 4 años de experiencia en desarrollo de software, construyendo aplicaciones web, plataformas digitales, APIs, servicios backend, bases de datos, despliegues y productos escalables. Combino desarrollo Full Stack con experiencia aplicada en IA, integrando LLMs, RAG, APIs de inferencia, routing por intención y pipelines backend en productos reales. Actualmente trabajo con soluciones de IA en producción, incluyendo serving de modelos LLM, optimización de latencia/throughput y agentes de IA.

HABILIDADES TÉCNICAS

- **Lenguajes de programación:** JavaScript, TypeScript, Python, Java, PHP, Bash
- **Frontend:** React.js, Next.js, Tailwind CSS, MUI, Shadcn/ui, Sass, Styled-Components, Bootstrap, Redux, Zustand, Context API, TanStack Query, Zod, Figma
- **Backend:** Node.js, Express.js, FastAPI, REST APIs, Socket.IO, Prisma ORM, Sequelize, Mongoose, Postman, Jest
- **Bases de Datos:** SQL, PostgreSQL, MongoDB, MySQL, Firebase (Realtime Database y Cloud Firestore)
- **DevOps y Metodologías:** Docker, AWS, Git, GitHub, Scrum, SEO, Strapi, WordPress
- **IA / ML Aplicada:** LLMs, Retrieval-Augmented Generation (RAG), FAISS, embeddings, PaddleOCR, Hugging Face, vLLM, TensorRT, NVIDIA Triton Inference Server, NVIDIA NIM, AI Agents, routing por intención, Integración de APIs
- **Desarrollo asistido por IA:** Cursor, Claude Code, Spec-Driven Development

EXPERIENCIA PROFESIONAL

IAMEX – Inteligencia Artificial México

Ciudad de México, México · Remoto

AI Software Engineer / Team Lead

Marzo 2025 – Presente

- Lidero optimización y serving de modelos LLM en producción sobre infraestructura NVIDIA A100/H100/B200, mejorando latencia, throughput y eficiencia operativa, incluyendo modelos OSS 20B y 120B, expuestos mediante APIs consumidas por clientes bajo un esquema de créditos, un chat de IA y asistentes personalizados basados en RAG.
- Dirigí técnicamente el desarrollo de un Chat de IA en producción e implementé un servicio de routing/orquestación por intención para seleccionar modelos, búsqueda en línea, generación de imágenes, generación documental y otros flujos según la intención del prompt del usuario.
- Participé en el desarrollo de una plataforma RAG para generar asistentes de IA personalizados por cliente, integrando carga de documentos, recuperación de contexto, subdominios dedicados y chats desplegados en producción que responden con información del negocio del cliente.
- Desarrollé y evalué un sistema RAG documental con ingesta, segmentación, embeddings, FAISS, recuperación semántica/híbrida, OCR para archivos escaneados con PaddleOCR, reranking, guardrails y validación de respuestas para mejorar precisión, trazabilidad y reducción de alucinaciones.
- Trabajo con Python, TypeScript, Node.js, Docker, Hugging Face, vLLM, TensorRT, NVIDIA Triton Inference Server y NVIDIA NIM para integrar software, modelos e infraestructura.
- Uso Cursor y Claude Code con Spec-Driven Development para definir especificaciones técnicas, planear tareas, implementar funcionalidades y revisar código.
- Ejecuté múltiples ciclos experimentales de fine-tuning con QLoRA sobre modelos como Mistral, DeepSeek, Qwen, Llama y Phi-3, evaluando viabilidad técnica antes de producción.

Towers Cem - Empresa Cementera

Barranquilla, Colombia · Remoto

Full Stack Developer / Backend Team Lead

Julio 2024 – Diciembre 2024

- Colaboré con un equipo internacional de desarrolladores de México, Colombia y Argentina en una plataforma digital para una empresa de cemento ecológico.
- Asumí liderazgo del equipo backend, supervisando calidad técnica, estructura de servicios, avance del desarrollo y coordinación de entregables.
- Participé en diseño UI con Figma, desarrollo de vistas con React.js, Next.js, Tailwind CSS y despliegues frontend/backend con Node.js, Express, PostgreSQL y AWS.

JMX – Agencia de Publicidad y Desarrollo Web

Full Stack Developer

Tlapa de Comonfort, Guerrero

Junio 2022 – Junio 2024

- Diseñé, desarrollé, desplegué y di mantenimiento de aplicaciones y sitios web, utilizando tecnologías de vanguardia, incluyendo Next.js, React.js, Node.js, Express, MongoDB, Docker, Tailwind CSS, SASS, Styled-Components, TypeScript, Jest, Git, GitHub y SEO para mejorar la indexación y velocidad de los sitios web.
- Realicé modelado de bases de datos, desarrollo de frontend y backend, integración de APIs, pruebas, mantenimiento y optimización SEO.

EXPERIENCIA ADICIONAL

Profesional independiente / Freelance

Remoto

Full Stack Developer, Contrato / Medio tiempo

Octubre 2024 – Noviembre 2025

- Desarrollé aplicaciones web para clientes de México y España utilizando React.js, Next.js, TypeScript, Node.js, Express, MongoDB, PostgreSQL, Docker y Tailwind CSS.

JMX – Agencia de Publicidad y Desarrollo Web

Tlapa de Comonfort, Guerrero

Desarrollador Frontend y Creador de Contenido

Enero 2018 – Junio 2022

- Diseñé, maqueté, administré y di mantenimiento de sitios web desarrollados en PHP con Joomla para diversos ayuntamientos y medios de comunicación.

Despertar de la Montaña – Medio de Comunicación

Tlapa de Comonfort, Guerrero

Frontend Engineer

Enero 2017 – Diciembre 2017

- Diseñé y maqueté su sitio web de noticias con HTML, CSS y JavaScript
- Diseñé, desarrollé, desplegué y di mantenimiento de una aplicación móvil de noticias con notificaciones push para Android e iOS, en Ionic con Angular.

Instituto Marcial de Taekwondo

Tlapa de Comonfort, Guerrero

Java Developer

Febrero 2016 – Diciembre 2016

- Desarrollé un sistema de puntuación para competencias de Taekwondo, con Java y MySQL. Mismo que fue implementado con éxito en varios torneos organizados por el mismo Instituto.

EDUCACIÓN

Desarrollo Web Full Stack y Ciencias de la Computación

Academlo (Enero 2024 – Julio 2024)

Ingeniería en Sistemas Computacionales

Instituto Tecnológico Superior de la Montaña (Agosto 2013 - Mayo de 2018)

HABILIDADES PROFESIONALES

Comunicación técnica, Colaboración remota, Liderazgo de equipo, Resolución de problemas, Pensamiento crítico, Gestión de prioridades y Aprendizaje continuo

CERTIFICACIONES

Certificado de Participación Destacada - Desarrollo Profesional

Institución: Academlo

Fecha de obtención: Octubre, 2024

Detalles: Reconocimiento por impartir talleres intensivos de más de 3 horas durante 5 semanas, de Desarrollo Full Stack con Next.js, Node.js, Typescript y MongoDB en el programa de egresados de Academlo.