

Angel Mauricio Marcos Montes

SENIOR SOFTWARE ENGINEER & AI ARCHITECT

contacto@angelm.dev
angelm.dev · github.com/Angel45604
linkedin.com/in/angelmmm · CDMX, México

Senior fullstack engineer with a product-first approach. Lately that means production LLM agent systems: a platform-wide voice and chat assistant, a conversational workflow builder, and the evaluation, cost, and review tooling that keeps them honest. Earlier years took me through payments, insurance platforms, and financial research tools. I like owning a problem all the way from requirements to production.

● EXPERIENCE

MindCloud

Aug 2025 – Aug 2026

Cloudy · Platform AI Assistant & Conversational Workflow Builder

- **Voice + chat, one conversation.** Hands-free realtime voice (OpenAI Realtime API over WebRTC) with model-decided turn-taking via semantic VAD and server-owned barge-in, sharing a single thread with typed chat, reachable from anywhere in the product.
- **Multi-model agent topology.** A low-latency voice orchestrator delegating to a Claude-powered insights sub-agent with a frozen, read-only tool surface, and a two-tier workflow builder that replaces manual drag-and-drop setup with conversation.
- **The provider key never reaches the browser.** Voice sessions are minted through a server-side relay that owns the key, the model choice, and the per-user permission context; the client only ever holds its own session.
- **"Navigate anywhere" + "Live Canvas."** Intent-based navigation over a permission-filtered, single-source route manifest with named-entity resolution, and a composable visual answer canvas rendered for whichever surface the conversation is on.
- **Reliable autonomous builds.** Task-boundary detection solving agent context loss, dynamic lookup resolution against live external APIs (Slack, Google Drive, Salesforce), and graph-corruption auto-repair.
- **Least-privilege security.** Cross-company access uses opaque short-TTL scope tokens re-verified on every use; database insight tools are read-only, admin-only, and deny by default.
- **Full-stack foundations.** SSE streaming, a tool-approval system with a 3-tier priority chain, conversation persistence, and a bulk AI-summary + PDF export pipeline on Redis background jobs.

Senior Software Engineer Oowlish

Mar 2025 – Aug 2025

- Built and maintained an end-to-end insurance platform used by agents and agencies nationwide; migrated it to an internal UI library, reducing code complexity and accelerating feature delivery by ~20%.
- Cut development time for new features by 2x by bringing AI coding tools (Claude Code, Cursor) into the team's workflow.

Senior Software Engineer EPAM Systems

Sep 2023 – Mar 2025

- **Hospitality client (B2C, millions of active users):** owned the Apple Pay integration from design through launch, reducing checkout time and improving conversion rates across platforms.
- **Software & high-tech client (R&D, internal tools):** built a surface-mock prototyping experience enabling designers to test new formats in days instead of weeks, backed by a flexible mock-data content service, plus reusable web components adopted by other teams.

Senior Software Engineer Ernst & Young GDS

Sep 2022 – Sep 2023

- Led a frontend team on a global financial research platform tightly integrated with Azure, used by thousands of analysts worldwide; CI/CD pipelines in Azure DevOps reduced release times by 40%.
- Integrated internal and external data services via Azure APIM, enabling scalable ingestion of high-value financial datasets.

Independent Technical Consultant

2018 – 2022

- Delivered custom web applications for SMB clients as a solo product owner, from requirements to deployment, replacing manual operations and cutting administrative workload by 30–50%.
- **DGETI (2020):** full-stack enterprise app digitizing institutional procedures with role-based access control and multi-step auditing; ETL consolidating Excel and database data reduced manual work by 70%, with on-demand reporting and full audit trails.

● SKILLS

AI & agents

LLM agent architecture, multi-agent orchestration, OpenAI Realtime API, Anthropic Claude, MCP, RAG, tool-calling, agent evaluation, prompt caching, LLM cost & observability, WebRTC voice, semantic VAD

Engineering

JavaScript, Node.js, Python, Swift, SwiftUI, React, Angular, RESTful APIs, SSE, ETL, RBAC

Cloud & infra

AWS, Azure (DevOps, APIM), GCP, Firebase, Redis, CI/CD

● OPEN SOURCE

the-foreman

Governance tooling for autonomous coding agents, shipped as a Claude Code plugin. A declarative lifecycle contract stops the agent at five hard human sign-off points, and a cross-vendor reviewer puts every Claude-authored change in front of an OpenAI model running sandboxed with zero write access: approval is refused outright if even one changed file went unreviewed.

● EDUCATION

Instituto Politécnico Nacional

B.Eng. studies, Bionic Engineering (not completed)

Universidad Autónoma Metropolitana

B.Eng. studies, Computer Systems Engineering (not completed)

Platzi

Web Development · 2017 – present

● LANGUAGES

English (advanced) · Spanish (native)