SENIOR AI ENGINEER · BEVERLY HILLS, CA · REMOTE (US HOURS)

Vladi Mijatovic

Senior AI Engineer | Agentic LLM Systems, Claude Code and Codex Tooling, MCP Servers, AI Voice Agents

ai@vladii.com | vladii.com | Beverly Hills, CA | Remote (US business hours)

100+	30+	20+	10+
CLAUDE CODE SKILLS AND COMMANDS	PROJECTS	APPS AND SITES LIVE	CUSTOM MCP SERVERS

01 PROFESSIONAL SUMMARY

Senior AI engineer who ships production AI agents end to end, from the LLM toolchain down to billing and uptime. Builds multi-tenant agentic LLM platforms in TypeScript and Bun, and a library of 100+ Claude Code and Codex skills, slash commands, lifecycle hooks, and custom subagents, plus multiple Model Context Protocol (MCP) servers on the official SDK. Deep on the Anthropic and OpenAI APIs with prompt caching, structured tool_use outputs, cost-aware multi-model routing, and RAG over encrypted knowledge graphs. Also builds conversational AI voice phone agents in English and Serbian and consumer apps live on the App Store and Google Play. **Builds the AI toolchain, not just uses it.**

02 CORE COMPETENCIES

Agentic LLM Systems · Claude Code and Codex Tooling · Custom MCP Servers · Multi-Model Routing · Prompt Caching and Cost Control · Structured Tool Use Outputs · RAG and Knowledge Graphs · AI Voice Agents · Full-Stack TypeScript · Supabase and Postgres RLS

03 WORK EXPERIENCE

Independent Senior AI Engineer / Founder

2023 – Present

Senior AI Engineer (AI Agents, Voice, Full-Stack and Mobile) · Remote (US hours)

- Built a multi-tenant agentic LLM platform in TypeScript and Bun (roughly 1,300 source and 570 test files) with a tool-calling agent loop, parallel sub-agent orchestration, per-tenant isolated state, a cost-aware router across Claude Opus, Sonnet, and Haiku with circuit-breaker fallbacks, prompt caching, and RAG over an encrypted knowledge graph.
- Authored 100+ reusable Claude Code and Codex skills and slash commands, lifecycle hooks, custom subagents, and multiple Model Context Protocol (MCP) servers on the official SDK, and ran senior multi-lens code review (correctness, security, performance, regression, UX) with cross-model adversarial passes before every ship.
- Designed and shipped production conversational AI voice phone agents on a managed platform (Retell, ElevenLabs neural TTS, Telnyx and Twilio telephony), provisioning the LLM, agent, and carrier number through REST and tuning turn-taking, interruption sensitivity, and voice pacing for natural human-paced calls, including a non-English agent hardened against dialect drift.
- Engineered LLM tool calling for live voice flows: calendar booking, qualified lead capture with a hotness score, human escalation, and warm transfer with caller-ID passthrough, plus a cron cost monitor that detaches the inbound number as a real kill switch on a spend breach, and fail-open per-caller and global rate limiting backed by the call ledger.
- Shipped and maintain a multi-tenant consumer AI app live on the App Store and Google Play (React Native, Expo SDK 54, React 19) on a Supabase Postgres backend with row-level security, released through EAS build and submit pipelines with over-the-air updates and Apple compliance handling, with billing across RevenueCat, native in-app purchases, and Stripe.

04 SELECTED PROJECTS

Custom AI Agent Toolchain (Claude Code, Codex, MCP)

100+ reusable Claude Code and Codex CLI skills and custom slash commands, plus multiple Model Context Protocol servers on the official SDK, including a Dockerized read-only connector secured with OAuth2 PKCE and JWT, and lifecycle hooks for secret-leak prevention, sacred-path edit guards, and git safety.

Model Context Protocol, @modelcontextprotocol/sdk, Hono, OAuth2 PKCE, JWT/jose, Docker, Anthropic Admin API, Bash, Python

Multi-Tenant Agentic Bot Fleet SELF-HOSTED, 24/7

Production agentic LLM platform (roughly 1,300 source and 570 test files) with a tool-calling agent loop, parallel sub-agent orchestration, per-tenant isolated state, cost-aware multi-model routing with circuit breakers, prompt caching, and RAG over an encrypted per-tenant knowledge graph; one isolated VPS per tenant with tiered watchdogs and synthetic liveness probes.

TypeScript, Bun, Anthropic SDK, Groq fallback, AES-256-GCM, pgvector, Hetzner Cloud, systemd, Sentry

Production AI Voice Phone Agent LIVE, 24/7

Conversational AI receptionist live on a carrier phone number, handling inbound calls with LLM-driven booking, lead capture, and human escalation. Built in two languages (including a Serbian variant with its own marketing site and callback demo) plus a from-scratch streaming voice agent over a WebSocket media relay with token streaming and mid-utterance barge-in, a daily-spend kill switch, and a multi-tenant call-analytics pipeline with idempotent webhook upserts.

Retell AI, custom-LLM websocket, ElevenLabs, Deepgram STT, Telnyx, Twilio ConversationRelay, GPT-4o/4.1, Cal.com API, Supabase, Postgres RLS, Vercel Cron

Multi-Tenant Production Consumer AI App IOS + ANDROID, LIVE

Live on the App Store and Google Play with an AI chat assistant, structured AI report generation across multiple product types, automatic PDF export, multi-language content, three subscription tiers, and a Supabase backend hardened with row-level security.

React Native, Expo SDK 54, React 19, TypeScript, Supabase, Postgres RLS, RevenueCat, Stripe, Anthropic Claude API, Sentry, PostHog, EAS Build

05 TECHNICAL SKILLS

AI / LLM and Agent Tooling: Claude Code (skills, slash commands, lifecycle hooks, subagents), Codex CLI, Model Context Protocol and custom MCP servers (official SDK), Anthropic Claude API (Opus, Sonnet, Haiku), OpenAI and Google Gemini APIs, prompt caching, structured tool_use outputs, cost-aware multi-model routing, circuit breakers, sub-agent orchestration, RAG, knowledge graphs, embeddings, pgvector, cross-model adversarial code review, prompt-injection hardening

Backend: Supabase, PostgreSQL, Row-Level Security, edge functions (Deno), realtime, Node.js, Bun, Python, REST APIs, webhooks, idempotency, rate limiting, server-side event tracking

Frontend / Mobile: React Native, Expo (SDK 53-55), React 19, Next.js 15/16 (App Router, SSR), Vite, TypeScript, Zustand, TanStack Query, Reanimated, Tailwind, shadcn/Radix, Zod, EAS Build, App Store Connect, Google Play Console

AI Voice / Telephony: Retell AI, custom-LLM websocket, ElevenLabs neural TTS (turbo v2.5, v3), Deepgram and Whisper-class STT, Telnyx, Twilio ConversationRelay, SIP termination, E.164, conversational system-prompt design, voice tool calling, multilingual voice (EN and Serbian sr-RS)

Infra / DevOps: Hetzner Cloud (hcloud), Linux (systemd, journald), Docker, Vercel, GitHub Actions, gitleaks, secret hygiene, Sentry, PostHog, synthetic probes and tiered watchdogs, Playwright automation

Payments / Monetization: Stripe (checkout, subscriptions, customer portal, webhooks), RevenueCat, native in-app purchases, server-side Apple and Google receipt verification, idempotent entitlement webhooks, double-charge concurrency guards

Languages: TypeScript, JavaScript, Python, Bash, SQL, Go (familiar)

06 EDUCATION

Master of Music, Jazz Piano and Composition · University of Music and Performing Arts, Munich (full scholarship) 2012

Bachelor of Music, Jazz Piano and Composition · University of Music and Performing Arts, Munich (full scholarship) 2010

Continuous self-directed AI/ML and production engineering · Self-directed 2023 - Present

07 BEYOND ENGINEERING

Award-winning jazz pianist and film composer in an earlier career. Master's and Bachelor's in Jazz Piano and Composition from the University of Music and Performing Arts Munich, both on full scholarship; named Best Jazz Musician in Europe; performed at Carnegie Hall; scored the film The Companion, winner of three Los Angeles Film Awards in 2024. Holds US citizenship through the EB-1A visa for extraordinary ability, and decades of performing under deadline built the discipline and attention to detail that now go into shipping production software.