

USAMA KHALIL

Software Architect | Principal Engineer | Full-Stack & Agentic AI | Azure Solutions Architect Expert

usamabinkhalil@gmail.com · UAE +971 58 829 4334 · Pakistan +92 336 404 2263

usamakhalil.com · [LinkedIn](#) · [GitHub](#) · UAE Resident (no sponsorship required)

PROFILE

Software architect with 11+ years designing, building and operating full-stack platforms across agentic AI, real-estate CRM, fintech and analytics. Architect and lead engineer of Televox AI, a multi-tenant voice and agentic AI SaaS running in production at sub-second p50 response, designed, shipped and operated end to end, including customer discovery and support. Previously led back-end and front-end architecture and mentored engineering teams through seven years at Optimageeks, and delivered front-end work on a CFD and forex trading platform serving millions of traders. Works AI-first, using Claude Code daily and reviewing model-generated code before it ships. Microsoft Certified: Azure Solutions Architect Expert and Azure AI Apps and Agents Developer Associate. Delivery experience across the US, Australia, UAE and Pakistan.

SELECTED OUTCOMES

- p50 voice response held at 0.8-1.0s under load on a real-time agentic pipeline I designed (Televox AI).
- Five production services operated solo, with a stateless worker fleet, admission control and per-turn latency and cost instrumentation.
- Self-hosted LLM inference server built and operated on Apple silicon: 70-84 tokens/s, seven days unattended uptime, a GPU memory failure diagnosed and resolved at 4x the context.
- Real-estate agencies on Optima CRM reported roughly a 40% revenue increase and the removal of manual listing errors.
- Front-end trading interfaces shipped for a global CFD and forex platform used by millions of traders.

TECHNICAL SKILLS

Architecture: Multi-tenant SaaS design and tenant isolation, microservices, real-time and event-driven systems, REST API design, database schema design, scalable system design, performance optimisation, observability and instrumentation

Languages & Frameworks: JavaScript, TypeScript, Python, PHP; Node.js (Express, NestJS, Sails, LoopBack), React, Next.js, React Native, Angular, FastAPI, Laravel, Yii2

AI Engineering: LLM integration (OpenAI, Anthropic, Gemini, Azure OpenAI) with tool calling and agent orchestration, prompt engineering with structured output enforcement, RAG with pgvector, Azure AI Foundry, AI evaluation and guardrails, cost and latency optimisation per conversation, NLP, sentiment analysis; self-hosted open-weight model inference (MLX on Apple silicon, quantised models, KV-cache and context budgeting)

Real-time & Integrations: WebRTC (LiveKit), WebSockets, SSE, real-time speech (Deepgram STT/TTS, Silero VAD), Twilio SIP and telephony, OAuth2/OIDC, signed webhooks, third-party API and CRM integration, n8n workflow automation

Data & Infrastructure: PostgreSQL (Supabase, pgvector), MySQL, MongoDB, Redis, AWS, Azure, Fly.io, Docker, CI/CD, Nginx, Apache, Socket.io, Git/GitHub; Tailscale private networking, macOS service administration (launchd, pmset, sysctl), performance benchmarking

AI-Assisted Development: Claude Code daily across design, implementation and testing; spec-driven workflows; reviewing and validating model-generated code for correctness, performance, tenant safety and idempotency

Leadership & Delivery: Technical direction and architecture ownership, code review and mentoring, customer discovery and direct stakeholder engagement, cross-functional delivery, Agile/Scrum, incident triage

CERTIFICATIONS

Microsoft Certified: Azure Solutions Architect Expert · Microsoft · September 2026 · [Verify credential](#)

Microsoft Certified: Azure AI Apps and Agents Developer Associate · Microsoft · September 2026 · [Verify credential](#)

Microsoft Certified: Azure Administrator Associate · Microsoft · September 2026 · [Verify credential](#)

EXPERIENCE

Architect & Lead Engineer | Televox AI

Late 2025 - Present · televox.ai · live paying customers

- Sole engineer on a multi-tenant agentic AI SaaS, architected and operated end to end: a single agent configuration answers inbound PSTN calls over a Twilio Elastic SIP trunk, web voice over LiveKit WebRTC, and an embeddable website chat widget.
- Designed the real-time pipeline (Silero VAD to Deepgram Nova-2 STT to OpenAI with tool calling to Deepgram Aura2 TTS) with barge-in and semantic turn detection, holding p50 response between 0.8 and 1.0 seconds measured under load.
- Own the production estate: five services, a stateless voice-worker fleet on Fly.io with LiveKit load balancing, admission control, and per-turn STT/LLM/TTS latency instrumentation surfaced through a latency-waterfall inspector.

- Built per-agent knowledge bases with pgvector RAG, HMAC-signed webhooks, real-time Cal.com booking, SMS actions and warm call transfers, plus outbound campaigns with TCPA safeguards.
- Run customer discovery and support directly: translate what small businesses describe into features, then ship, monitor and iterate on them. Self-serve telephony provisioning, plan enforcement and billing.

Principal App Developer | Medical Guardian

May 2026 - Present · United States · Remote

- Principal developer on React Native applications and the API integration layer for a US medical alert and connected-care provider.
- Define front-end structure and API contracts for new features with product, back-end and QA, within the reliability and compliance expectations of a healthcare product where device and response workflows cannot fail.

Senior Software Engineer | Pepperstone

Apr 2024 - May 2026 · Australia · Remote

- Front-end engineering on a global CFD and forex trading platform serving millions of traders, partnering with a C# back-end team to deliver trading features and conversion flows.
- Built React and TypeScript interfaces for the crypto trading experience, including real-time market data delivered over WebSockets.
- Implemented secure authentication meeting financial-industry compliance requirements.

Senior Software Developer | Modulus

Sep 2022 - Mar 2024 · United States · Remote

- Senior developer across the product portfolio, including Anzer, turning business requirements into system design and specifications.
- Designed and ran integration and performance testing across multiple environments to protect system stability.
- Coordinated priorities across cross-functional teams and led triage of development and production issues.

Software Engineer to Senior Software Engineer to JavaScript Full-Stack Developer | Optimageeks

Jan 2015 - Aug 2022 · Lahore, Pakistan

- Led back-end architecture for core web applications, covering database schema design, server-side API design and server provisioning.
- Led front-end architecture and designed the user-facing interactions and flows across client platforms.
- Mentored junior engineers through code review and pairing, and set the review standards the team worked to.
- Designed and built REST APIs for system integration and inter-service data exchange, and integrated internal applications with third-party systems. Principal client platforms: Optima CRM, My Lobby, LodgeIntel and 100Bet.

SELECTED PROJECTS

Televox AI · Multi-tenant voice and agentic AI SaaS · LiveKit, Twilio SIP, Deepgram, OpenAI, Supabase, Next.js, Python, Fly.io · Architect and lead engineer

Local LLM Inference Server · Private AI coding infrastructure · MLX, LM Studio daemon, Tailscale · Designed, built and operated

- Headless server on Apple silicon (M4 Pro, 24 GB) serving open-weight models through OpenAI and Anthropic-compatible APIs, published only on a private Tailscale network, with launchd bring-up, a persisted GPU wired-memory limit and a crash watchdog. Seven days unattended uptime, in daily use.
- Benchmarked at 70-84 tokens/s generation and ~450 tokens/s prompt processing; diagnosed a GPU out-of-memory failure from engine growth and resolved it with a 4x smaller KV-cache at a 64K context.

CallAgent AI · Real-time conversational AI for call centres · NestJS, React, Docker, Twilio, OpenAI, Gemini · Lead developer

- Led development of an AI calling platform automating inbound and outbound calls, with conditional real-time transfers, Cal.com booking, post-call SMS and email, automated information extraction, a campaigns module wired to Google Sheets and Zoho CRM, and an admin panel for prompts and knowledge bases.

Optima CRM · Real-estate CRM and command centre · MEAN stack · Full-stack developer

- Replaced fragmented manual processes with one platform: lead capture from portals and social channels, centralised listings with automated portal syncing, digital offers, invoicing, commission tracking, rentals and a real-time BI layer. Agencies reported roughly a 40% revenue increase and the end of manual listing errors.

Also delivered: My Lobby, visitor management with QR access and an AI call agent (MERN, React Native) · LodgeIntel, hospitality analytics (MERN) · PlanetPop, English-learning platform built in Oxford (Angular, Node.js) · The Good Inside, headless commerce (Next.js, WordPress) · 100Bet, iGaming redesign (Angular) · DG-Agency, portfolio platform (PHP/Yii2)

EDUCATION

BSc Electronics and Communication · University of Lahore · Sep 2008 - Mar 2015