Joaquim José da Fonseca Viana

Engenharia de software com foco em IA. Agentes, RAG, voz e mídia generativa em produção.

Brasília, DF jjfvwork@gmail.com +55 61 99515-1699 linkedin.com/in/joaquim-josé-da-fonseca-viana
github.com/jj-viana joaquim.eipe.com.br

Construo sistemas de IA em produção há dois anos: sistemas multiagente, RAG com avaliação, agentes de voz em tempo real, pipelines de mídia generativa em GPU e servidores MCP. Decido modelo e arquitetura com avaliação (A/B com LLM como juiz, suítes contra modelos reais, benchmarks de modelos locais) e trato custo de LLM como requisito de produto. Palestrante sobre sistemas agênticos na UnB e no MVP Conf 2025.

Experiência

CrazyTechLabs, time de IA e Inovação nov/2024 – atual

Desenvolvedor de software desde dez/2025; estagiário de nov/2024 a dez/2025.

- Sistema multiagente de criação de conteúdo em produção (Google ADK, Gemini, Azure Durable Functions): coordenador e especialistas de pesquisa web, roteiro, imagem e vídeo. A migração de arquitetura foi decidida por A/B com usuário simulado e LLM como juiz.
- Maior contribuidor da API (711 de 1.582 commits) e do app web de uma plataforma SaaS multi-tenant de redes sociais (.NET 9, Next.js 16, Azure Functions, Service Bus): publicação e agendamento em Instagram, Facebook, LinkedIn e X, campanhas, respostas automáticas e custo de LLM por cliente via LiteLLM.
- Assistente que opera a própria interface (Python, FastAPI, SSE): guia o próximo passo, clica e preenche na tela ou executa pela API, com 14 ferramentas, permissões do usuário e confirmação de ações irreversíveis. 86% dos commits. Agente de voz em tempo real para o onboarding.
- Infraestrutura de GPU compartilhada sob demanda e servidor MCP para pipelines ComfyUI; estúdio de TTS com controle de qualidade automático (WER, similaridade de voz, MOS previsto).

ForgeX Core, fundador 2026 – atual

- Sccela (sccela.com.br): app de finanças com assistente de ~31 ferramentas, ações com confirmação humana, RAG em pgvector e leitura de faturas e notinhas. Suítes de avaliação contra modelos reais guiam o roteamento de modelo por custo. 689 commits, lançado em out/2026.
- Plataforma acadêmica multi-instituição: schema por instituição, identidade plugável (Google, Entra ID) e 7 ADRs.

Faculdade Brasília, tecnologia ago/2022 – atual

- Sistema acadêmico em produção (NestJS, Next.js, PostgreSQL): do plano de ensino ao curso montado no Moodle por um plugin PHP próprio; sincronização de matrículas com travas de segurança e postmortem de incidente.
- Site, vestibular online, pagamentos Pix e medição de conversão (Meta CAPI, GA4, LGPD). Auditoria de R$208 mil em anúncios com o servidor MCP da Meta.

Projetos pessoais

Tutor de idiomas por voz. Avatar 2D e 3D com lip-sync e quadro de aula sincronizado com a fala, memória de longo prazo consolidada por LLM, três motores de voz.

Estudo com citações verificadas. RAG híbrido com RRF e rerank; avaliação de 3 modelos em acerto, custo e latência p95.

Monitor de passagens. Agente com MCP próprio, sete fontes de dados e confirmação humana.

Vídeo narrado 100% local. LLM, difusão, vídeo e voz clonada numa GPU de 10 GB, com QA por WER.

Palestras

Fundamentos de IA e sistemas agênticos. SEMUNI, UnB.

MVP Conf 2025. Protótipos no-code, desenvolvimento e Azure.

Formação

Engenharia de Software, Universidade de Brasília. mar/2024 – dez/2028 (previsão).

Inglês B2. Exame Cambridge; curso completo no Thomas Jefferson.

Certificações

Azure Cosmos DB AI Developer Associate (DP-420), Microsoft, 2026

Agentspace (Deploy e Extend) e Workspace Deployment Services, Google, 2025

Fortinet Certified Fundamentals in Cybersecurity, 2024

Competências

Linguagens: Python, TypeScript, C#, SQL, C.

IA: OpenAI e Azure OpenAI, Gemini, Claude, Ollama; Google ADK, Vercel AI SDK, MCP, LiteLLM; pgvector; ComfyUI, TTS e STT; LLM como juiz.

Plataforma: Azure, Google Cloud, Docker, GitHub Actions, Bicep, PostgreSQL, Cosmos DB.

Como trabalho: agentes de código (Claude Code), especificação antes do código, testes e CI.