

KELVIN BIFFI

Engenheiro de Software Sênior | IA Aplicada, RAG e Avaliação de LLM

Balneário Camboriú, SC (UTC-3) · Remoto · kelvinbiffi.developer@gmail.com · linkedin.com/in/kelvinbiffi · kelvinbiffi.com

RESUMO

Engenheiro de software com mais de 15 anos entregando sistema em produção, boa parte disso em software que carrega dinheiro e identidade: checkout de alto volume para rede nacional de farmácia, gateway de autenticação em Keycloak com MFA para milhões de usuários, cobrança de assinatura em Stripe, e plataforma americana de dado de saúde com controle de acesso por cliente.

Nos últimos anos essa engenharia foi apontada para sistema de LLM: agentes com tool calling, servidores MCP, pipeline de recuperação, e a camada de validação que impede o modelo de mentir. **Agente não falha alto.** Ele devolve algo plausível, o log fica limpo, e semanas depois alguém encontra um registro corrompido. Essa camada de confiabilidade é a mesma disciplina que checkout e autenticação já exigiam.

O trabalho com dado e linguagem é mais antigo que a onda de LLM: fiz **análise semântica e de sentimento** de texto aberto de cliente para decidir feature de produto antes de existir GPT em produção.

EXPERIÊNCIA

AI Automation Engineer, Hire With Near (plataforma americana de recrutamento) · 2026

- Construí o **Near Companion** ponta a ponta em Python e TypeScript: extensão de Chrome mais pipeline de LLM sobre Recall.ai, Supabase, HubSpot e n8n. Um **bot entra na call como participante silencioso**, gera a descrição de vaga estruturada ao vivo e alimenta o matching de candidato na mesma sessão.
- Toda saída de modelo passa por validação de schema e asserção campo a campo antes de poder tocar um registro de CRM.** Chamada de API bem-sucedida não diz nada sobre o payload fazer sentido, então campo obrigatório, faixa de valor e pertinência de enum são checados separado do parse.
- Falha é explícita, nunca silenciosa.** Saída inválida é repetida com prompt corretivo e, se insistir, cai para rascunho sinalizado para revisão humana, em vez de gravar dado parcial.
- Descoberta, matching e ranqueamento de candidato com recuperação contextual, apresentando e **explicando** por que cada candidato é relevante.

AI Workflow Engineer, 8 Figure Agency (EUA) · 2026

- Sistemas multiagente com integração de ferramenta e API, usando **orquestração explícita em vez de framework**, para que toda chamada de ferramenta fique inspecionável e reproduzível.
- Guardrail em ação irreversível:** nenhum agente grava em CRM, dispara contato ou gasta verba sem passar por validação, e o que fica abaixo do limiar para humano.
- Pipeline de Meta Ads para BigQuery ponta a ponta (API, backfill, alinhamento de schema contra dado de custo no Airtable) e pipeline de lead em n8n sobre Gmail, GoHighLevel e Apollo.

Engenheiro de Software Sênior, Firstleaf (e-commerce de assinatura, EUA) · 2021-2026

- Dono de assinatura, cobrança e retenção em **Stripe com ciclo de vida orientado a webhook**, onde webhook perdido ou duplicado é incidente de cobrança, não relatório de bug. Sistema de landing e experimentação A/B em escala, GraphQL via Gatsby, integração com Shopify.
- Plataforma de conteúdo de SEO com LLM:** geração programática de conteúdo de produto e categoria em escala de catálogo, ancorada em dado estruturado de produto e **validada antes de publicar**, para que o texto gerado não pudesse afirmar atributo que o produto não tem. Nesse volume ninguém lê toda página, então a correção tem que ser imposta pelo pipeline.
- Análise semântica e de sentimento de feedback de cliente:** review de texto aberto e mensagem de suporte classificados e agrupados em tema e sentimento, transformando texto não estruturado em insumo de decisão de produto. Isso rodou antes da onda de LLM, e é por isso que minha exposição a ML é mais antiga que meu trabalho com LLM.

Tech Lead, Zallpy / Surgo Health (plataforma americana de dado de saúde) · 2024-2025

- Plataforma de pesquisa e análise de dado de saúde vendida para cliente farmacêutico, em **Next.js, Express e PostgreSQL**. Liderei o lado de engenharia da entrega com o cliente americano.
- Construí **níveis de acesso e de granularidade de dado por cliente**, de forma que cada cliente vê apenas a resolução de dado de saúde que o contrato dele permite. **O modelo de acesso é o produto, não uma configuração.**
- Camada geográfica em **Leaflet** combinando dado de saúde com mobilidade e dificuldade de acesso a atendimento, para que a análise levasse em conta quem consegue fisicamente chegar ao cuidado e quem não consegue.

Engenheiro de Software Sênior, FCamara / Rede Drogasil · 2021-2025

- **Migração de Magento para Next.js da vitrine inteira, rota por rota, sem janela para derrubar o checkout.** Primeiro na squad de migração, depois nas squads de jornada do cliente e de checkout.
- Dono da jornada ponta a ponta (listagem, carrinho, checkout, confirmação de pedido) em alto volume de pedido, com serviços em **RabbitMQ**, Apollo Client em produção, cache e otimização de query.

Engenheiro de Software Sênior, Zartis (clientes corporativos, EUA e Dubai) · 2021-2024

- **Knowledge Gap, LLM sobre documentação de cliente.** O cliente conecta a própria plataforma de documentação, o sistema lê o corpus e aponta **lacuna e contradição**: o que falta, o que dois documentos afirmam de forma diferente, o que está desatualizado. Recuperação sobre o corpus do cliente com análise de LLM em cima. **LLM corporativo entregue a cliente pagante, não protótipo.**
- Frontend para o **DIFC, centro financeiro internacional de Dubai**, interface de grau financeiro onde a correção do que é exibido é questão regulatória e não preferência de UX, e para a **605 Media Analytics (EUA)**, interface analítica sobre grande volume de dado de mídia. Trabalhei direto com time de engenharia do cliente, em inglês, entre fusos dos EUA e dos Emirados.
- Engenharia de produto para cliente de **SaaS de real estate nos EUA**, no mesmo modelo de alocação.

Engenheiro Frontend Sênior, Zallpy / Lojas Renner · 2022-2023

Liderei o sistema corporativo de login e registro de uma rede nacional de varejo. **Keycloak com MFA, gateway de autenticação para milhões de usuários.**

Consultoria e Engenharia de Produto · 2011-2021

Pmweb: sistema de landing e CRO, incluindo um **construtor de landing page em Vue.js que reduziu o esforço de desenvolvimento em 70%**. Também K3 Company, OSF, Object Edge e Unilasalle: plataforma de e-commerce, otimização de receita em AdTech, ferramenta de automação.

PROJETOS PRÓPRIOS

Projetos que eu construo e mantenho. Não é trabalho de cliente.

- **Mentezinha.app**, IA conversacional child-safe em produção no Google Vertex AI, e o sistema de LLM que eu consigo percorrer ponta a ponta. **Recuperação por criança em busca vetorial gerenciada**: cada criança tem um índice que cresce além do que o pai configurou, acumulando entendimento ao longo das conversas. O guardrail é uma **camada de validação com autoridade para recusar**, não uma instrução no prompt, porque com criança como usuária a recusa é muitas vezes a resposta correta.
 - **Ferramenta própria de medição de desperdício de contexto em agente**: lê o trace de execução e atribui o gasto a causa específica, como schema de ferramenta enviado em toda chamada e nunca invocado, payload cru de ferramenta no contexto, chunk quase duplicado no retrieval, e retry repagando chamada já concluída. **Detectores determinísticos, sem modelo no circuito**, para o resultado ser reproduzível.
 - **Portão determinístico de número**: um post gerado citou um fato válido, passou no portão de rastreamento, e ainda assim inventou um número que não existia na fonte. A validação ficou verde. **Rastreabilidade não é verificação**, então passei a comparar todo número da saída contra o corpus de origem.
 - **StartAtende**, suporte por WhatsApp com IA, roteamento de intenção, custo por conversa e análise de sentimento sobre as conversas do próprio agente. · **Linkezera.com**, SaaS em produção com mais de 3.700 usuários. · **MentalKare** e **AdsOptimizer**. · Servidores MCP, integração multiagente, pipeline de mídia com FFmpeg e WhisperX.
-

STACK

Linguagens Python, TypeScript, JavaScript, SQL, PHP · **IA e LLM** OpenAI API, Gemini API, Vertex AI, agentes, multiagente, tool e function calling, **MCP (servidor e cliente)**, prompt e context engineering, RAG e recuperação, busca vetorial gerenciada, embeddings, guardrails, **validação e avaliação de LLM**, NLP aplicado, análise de sentimento, WhisperX · **Backend** Node.js, Express, FastAPI, REST, GraphQL, Apollo, webhooks, RabbitMQ, Stripe · **Frontend** React, Next.js, Gatsby, extensão de Chrome, TailwindCSS · **Dados** PostgreSQL, MySQL, MongoDB, Supabase, BigQuery · **Auth e Plataformas** Keycloak, MFA, HubSpot, Shopify, Recall.ai · **Teste e Operação** Vitest, Cypress, Playwright, Docker, CI/CD, GitHub Actions, Sentry, CloudWatch

FORMAÇÃO

MBA em Inteligência Artificial, *em andamento* · MBA em Market Intelligence · Pós-graduação em Liderança e Coaching · Bacharelado em Análise e Desenvolvimento de Sistemas

Autor de quase trinta livros técnicos publicados, quatro deles exatamente no stack em que trabalho: *MCP na Prática com Python*, *Context Engineering: O Guia Definitivo para LLMs*, *Agentes de IA com Python* e *Spec Driven Development na Prática*.

Idiomas: Português (nativo), Inglês (fluyente, proficiência profissional)