

KELVIN BIFFI

Senior Software Engineer | AI Systems & Product Engineering

Brazil (UTC-3) · Remote only · Full-time kelvinbiffi.developer@gmail.com · linkedin.com/in/kelvinbiffi · kelvinbiffi.com

SUMMARY

Software engineer with 15+ years shipping production systems, most of it on software that carries money and identity: high-volume checkout for a national pharmacy chain, an authentication gateway on Keycloak with MFA serving millions of users, subscription billing on Stripe, and a US healthcare data platform with per-client access control.

For the last few years that engineering has been pointed at LLM systems: agents with tool calling, MCP servers, retrieval pipelines, and the validation layer that keeps them from lying. An agent does not fail loudly. It returns something plausible, the logs look clean, and weeks later someone finds a corrupted record. That reliability layer is the same discipline checkout and authentication already demanded. Python, TypeScript, PostgreSQL, React and Next.js, delivered remotely for US, UK and UAE clients.

EXPERIENCE

AI Automation Engineer - Hire With Near (US recruiting platform) · 2026

- Built **Near Companion** end to end in Python and TypeScript: Chrome extension plus LLM pipeline over Recall.ai, Supabase, HubSpot and n8n. A **bot joins the intake call as a silent participant**, generating the structured job description live and feeding candidate matching from the same session.
- **Every model output passes schema validation and field-level assertions before it can touch a CRM record.** A successful API call says nothing about whether the payload made sense, so required fields, value ranges and enum membership are asserted separately from the parse.
- **Failure is explicit, never silent.** Invalid output is retried with a corrective prompt, then falls back to a flagged draft for human review instead of writing partial data into the CRM.
- Candidate discovery, matching and ranking with contextual retrieval, surfacing and *explaining* relevant candidates in seconds.

AI Workflow Engineer - 8 Figure Agency (US) · 2026

- Multi-agent systems with tool and API integrations, using **explicit orchestration rather than a framework**, so every tool call stays inspectable and reproducible.
- **Guardrails on irreversible actions:** no agent writes to a CRM, sends outreach or spends budget without passing validation first, and anything below threshold stops for a human.
- Meta Ads to BigQuery pipeline end to end (API, backfill, schema alignment with Airtable COGS data) and n8n lead pipelines across Gmail, GoHighLevel and Apollo.

Senior Software Engineer - Firstleaf (US wine subscription e-commerce) · 2021-2026

- Owned subscription, billing and retention on **Stripe with webhook-driven lifecycle events**, where a missed or duplicated webhook is a charging incident, not a bug report. Landing systems and A/B experimentation at scale, GraphQL through Gatsby, Shopify integration.
- **LLM-driven SEO content platform:** programmatic generation of catalog-scale product and category content, grounded in structured product data and validated before publish, so generated copy could not assert attributes a product did not have. At that volume nobody reads every page, so correctness has to be enforced by the pipeline.
- **Semantic and sentiment analysis of customer feedback:** open-text reviews and support messages classified and clustered into themes and sentiment, turning unstructured customer text into an input for product decisions. This ran before the LLM wave, which is why my ML exposure predates my LLM work.

Tech Lead - Zallpy / Surgo Health (US healthcare data platform) · 2024-2025

- Health data research and analytics platform sold to pharmaceutical clients, in **Next.js, Express and PostgreSQL**. Led the engineering side of delivery with the US client.

- Built **per-client access and data granularity levels**, so each client sees only the resolution of health data their agreement allows. The access model is the product, not a setting.
- Geographic layer in **Leaflet** combining health data with mobility and access-to-care difficulty, so analysis could account for who can physically reach care and who cannot.

Senior Software Engineer - FCamara / Rede Drogasil (national pharmacy chain) · 2021-2025

- **Magento to Next.js migration of the full storefront, route by route, with no window to take checkout down.** Migration squad first, then the customer journey and checkout squads.
- Owned the journey end to end (product listing, cart, checkout, order confirmation) at high order volume, with **RabbitMQ** services, Apollo Client in production, caching and query optimization.

Senior Software Engineer - Zartis (US & Dubai enterprise clients) · 2021-2024

- **Knowledge Gap, LLM over client documentation.** The client connects their own documentation platform, the system reads the corpus and surfaces **gaps and contradictions**: what is missing, what two documents state differently, what is stale. Retrieval over the client corpus with LLM analysis on top. Enterprise LLM delivered to a paying client, not a prototype.
- Frontend systems for **DIFC, the Dubai International Financial Centre**, financial-grade interfaces where the correctness of what is displayed is a regulatory concern rather than a UX preference, and for **605 Media Analytics (US)**, analytics interfaces over large media datasets. Worked directly with client engineering teams in English, across US and UAE time zones.
- Product engineering for a **US real estate SaaS** client, building features across the platform in the same allocation model.

Senior Frontend Engineer - Zalpy / Lojas Renner · 2022-2023 Led the enterprise login and registration system for a national retail chain. **Keycloak with MFA, an authentication gateway serving millions of users.**

Consulting & Product Engineering · 2011-2021 Pmweb: landing systems and CRO, including a **Vue.js landing page builder that cut development effort by 70%**. Also K3 Company, OSF, Object Edge and Unilasalle: e-commerce platforms, AdTech revenue optimization, automation tooling.

PERSONAL PROJECTS

Side projects I build and maintain myself. Not client work.

- **Mentezinha.app** - child-safe conversational AI in production on Google Vertex AI, and the LLM system I can walk through end to end. **Retrieval per child on managed vector search**: each child has an index that grows beyond what the parent configured, accumulating understanding across conversations. The guardrail is a **validation layer with the authority to refuse**, not an instruction in a prompt, because with children as the users a refusal is often the correct answer.
- **StartAtende** - AI WhatsApp support with intent routing, per-conversation cost tracking and sentiment analysis over the agent's own conversations. · **Linkezero.com** - production SaaS, 3,700+ users. · **MentalKare** and **AdsOptimizer**. · MCP servers and tooling, multi-agent integrations, media pipelines with FFmpeg and WhisperX.

STACK

Languages - Python, TypeScript, JavaScript, SQL, PHP **Backend** - Node.js, Express, FastAPI, REST, GraphQL, Apollo, webhooks, RabbitMQ, Stripe **Frontend** - React, Next.js, Gatsby, Chrome Extensions, TailwindCSS **Data** - PostgreSQL, MySQL, MongoDB, Supabase, BigQuery **AI/LLM** - OpenAI API, Gemini API, Vertex AI, agents, multi-agent, tool/function calling, **MCP (servers and clients)**, prompt & context engineering, RAG and retrieval, managed vector search, guardrails, LLM validation & evaluation, WhisperX **Auth & Platforms** - Keycloak, MFA, HubSpot, Shopify, Recall.ai **Testing & Ops** - Vitest, Cypress, Playwright, Docker, CI/CD, GitHub Actions, CloudWatch

EDUCATION

MBA, Artificial Intelligence - *in progress* · MBA, Market Intelligence Postgraduate, Leadership & Coaching · BSc, Systems Analysis & Development

Author, almost thirty published technical books, four in the exact stack I work in: *MCP na Prática com Python*, *Context Engineering: O Guia Definitivo para LLMs*, *Agentes de IA com Python* and *Spec Driven Development na Prática*.

Languages: Portuguese (native), English (fluent - professional working proficiency)