

# Gustavo Berny

## Senior Software Engineer · Applied AI and Agents

AI Agents · RAG/pgvector · Claude Code · Node.js/TypeScript · PHP/Laravel · Python · PostgreSQL · React

+55 51 98151-8112 | [pelotas.berny93@gmail.com](mailto:pelotas.berny93@gmail.com) | [linkedin.com/in/gustavo-berny](https://linkedin.com/in/gustavo-berny) | [gustavoberny.com](https://gustavoberny.com)

## SUMMARY

Software engineer with 10+ years of experience, working full-stack on a solid backend foundation in **Node.js/TypeScript** and **PHP/Laravel (MVC)**. In recent years my work has centered on **applied AI in product**: I am the lead developer and architect of a multi-tenant SaaS platform of **AI agents** running in production — a **RAG** pipeline on **pgvector**, asynchronous agent orchestration, **tool calling** and **MCP**, with guardrails, observability and per-tenant cost control. The platform handles **~153k messages/month**, with **85% of conversations resolved without human intervention**. I also use coding agents as a core part of my engineering process.

I work at the intersection of engineering and product thinking — with prior experience as a **Product Owner**, I approach architecture with the user and the business outcome in mind. That lets me judge where AI creates real value and where it is just noise, delivering solutions that hold up in production across operations, cost and quality.

## CORE SKILLS

- **AI & Agents**: production AI agents, RAG with pgvector, multi-agent orchestration, tool/function calling, MCP (client and server), guardrails, multimodal AI, LLMops/FinOps.
- **Backend & Architecture**: Node.js/TypeScript (ESM), Express, Prisma, PHP/Laravel (MVC), scalable RESTful APIs, multi-tenant architectures, RabbitMQ queues and real-time with Socket.io.
- **Automation & Integrations**: Python for process automation and data processing; integrations with WhatsApp (Z-API and Meta Cloud API), external APIs and third-party systems.
- **Databases**: data modeling and SQL in PostgreSQL and MySQL, vector search with pgvector, performance tuning with functions, triggers and CTEs.
- **Frontend**: React, Vite, Tailwind, Redux Toolkit, in-house design system, JavaScript (ES6+), HTML5 and CSS3.
- **Quality & DevOps**: Vitest/Jest, unit, integration and E2E testing with Testcontainers, property-based testing, Docker, CI/CD, Sentry and Prometheus metrics.
- **Security**: JWT, per-tenant API keys, encrypted credentials, rate limiting, SSRF protection and audit trails.

## FEATURED PROJECT

### Agents-IA — Multi-Tenant AI Agent SaaS Platform · Imply

03/2026 – Present

A multi-tenant customer engagement platform where companies configure AI agents that talk autonomously to leads and customers, handing off to a human agent (**human-in-the-loop**) when needed. Channels: WhatsApp, embeddable web widget, email, public API and MCP. I served as lead developer and architect, with over **2,400 commits** and end-to-end ownership — from AI architecture to production infrastructure. Live with 11 provisioned tenants (7 active), 113 operators and 126 configured AI agents.

- **Delivered 85% of conversations resolved end-to-end by AI**, with no human intervention (6,570 of 7,725 closed in Aug/2026; 83.3% over the last 90 days), running ~153k messages and ~7.8k conversations per month — 24× growth in 4 months.
- **Cut time to first response from 2m39s to 4.7 seconds** (median, -97%), measured across 61k question-answer pairs in production.
- **Architected a multi-provider LLM layer** (OpenAI, Anthropic and OpenWebUI) behind a common interface, with retries, per-model capability detection and provider switching without touching business code.
- **Built an in-house RAG pipeline**: custom chunking, embeddings, vector search on PostgreSQL 16 with pgvector, query embedding cache, versioning and knowledge base access logging.
- **Developed an asynchronous agent orchestrator** with a 14-stage pipeline consumed from a RabbitMQ queue — agent loading, RAG retrieval, LLM call, tool execution, guardrails and delivery.
- **Implemented tool calling** with a tool registry, HTTP execution with encrypted credentials, preview, tool memory and per-call logging; plus **MCP (client and server)** with a server catalog, tool adapter and URL/SSRF guard.
- **Designed intelligent agent routing** with 7 priority levels (active conversation, routing rules, interactive menu, default agent, autoAssume), plus an intent router and message debouncer.
- **Applied production guardrails** (content filtering, blocking of unauthorized URLs, internal ID leak protection and tool argument validation) and multimodal features: audio transcription, vision, PDF/DOCX extraction, translation and language detection.

- **Instrumented LLMops/FinOps:** token consumption and cost calculation per tenant and per model, with a versioned pricing table and usage dashboards, keeping inference cost at \$0.12 per completed conversation (810 million tokens/month); plus sentiment analysis, conversation classification and SLA metrics.
- **Sustained quality at scale:** a ~107k-line monorepo across 1,019 files, 69 Prisma models and 109 migrations, ~3,250 test cases (Vitest, Testcontainers, property-based testing) with an 80% coverage threshold and an architecture check running in CI.
- **Stack:** Node 22, TypeScript, Express 5, Prisma 6, Zod, React 19, Vite 6, Tailwind 4, PostgreSQL 16 + pgvector, Redis, RabbitMQ, Socket.io, Docker, Bitbucket Pipelines, Oracle Cloud (OCI), Sentry and Prometheus.

## PROFESSIONAL EXPERIENCE

---

### Senior Backend Engineer & Project Lead — Imply

02/2025 – Present

- **AI Agent Platform (Agents-IA):** technical leadership and end-to-end development of the platform detailed above — AI architecture, backend, frontend and infrastructure.
- **Technical Leadership:** backlog refinement, definition of architecture standards and team alignment, making sure everyone understands what needs to be built and why.
- **AI-Assisted Development:** use of Claude Code with parallel agents in isolated git worktrees, sharply reducing delivery time without giving up review and quality.

### Full-Stack Developer — BernyFlow (Personal SaaS Project)

02/2025 – Present

- **Business Management SaaS:** full-stack development covering CRM, invoicing and financial reporting. Docker infrastructure, JWT authentication, REST APIs and a finance module with NFC-e electronic invoice issuing.
- **AI-Accelerated Delivery:** from zero to production on a Linux VPS, with over 300 deploys in 3 months working solo — AI agents assisting across the entire cycle.
- **Stack:** Node.js, TypeScript, React, PostgreSQL, Prisma, Docker — [bernyflow.com.br](https://bernyflow.com.br)

### Full-Stack Developer — Liga dos Vales Volleyball League (Personal Project)

2026 – Present

- **Official league portal:** built and maintain the public site for the largest volleyball championship in Santa Cruz do Sul, Brazil — stages, brackets, rankings, player transfer market, Hall of Fame and a news portal, with a restricted admin area for the organizers.
- **Live operation:** 2026 season with men's and women's stages gathering 20+ teams, publishing results and standings after every stage.
- **Stack:** Next.js, React, Docker on a self-managed VPS — [ligadosvales.com.br](https://ligadosvales.com.br)

### Mid-Level Full-Stack Developer — Sellflux

09/2023 – 01/2025

- **AI Chatbot:** built a high-performance chatbot, integrating vector search in PostgreSQL to improve answer relevance.
- **API & Testing:** backend with Node.js and Express, ensuring quality and reliability with Jest.
- **Databases:** performance tuning through advanced use of functions, triggers and CTEs.
- **Frontend:** complete, interactive interface built from scratch with React.

### Backend Developer — Imply

08/2021 – 09/2023

- **Billing System (PHP/Laravel):** developed a complete membership billing system, adapted for football clubs, using PHP in the MVC pattern.
- **Financial Transactions:** architected and oversaw a direct debit system for 107,000+ members, processing over R\$4 million (Brazilian reais) in monthly transactions.
- **Integrations & Automation:** solutions for cash flow control, document printing and access and billing rules, using PHP, JavaScript and XML.
- **Access Control:** parking management system processing over 4,000 vehicles per day.

### Technical Support Analyst — RM Software

02/2015 – 04/2020

- **ERP Systems:** support, implementation and user training on systems for account management, invoicing, tax documents and tax calculation, proposing improvements based on user feedback.

## EDUCATION

---

- **BSc in Data Science** — Unopar (in progress)
- **Technical Degree in Information Technology** — UNISC, Santa Cruz do Sul, Brazil

## LANGUAGES

---

- **English:** Advanced — fluent communication in professional and technical settings. | **Portuguese:** Native.