

Julio Gutierrez



✉ julioguti.dev | ✉ cv@julioguti.dev | in julio-guti | 🐧 bubuntux | 🚀 bubuntux

Staff Software Engineer · LLM inference platforms · 15+ years · 22M+ OSS downloads · Remote

Professional Experience

Staff Software Engineer

2021 - Present

Austin, TX | Remote

- Designed and delivered the company's first self-hosted SLM inference platform, taking it from architecture review to serving multiple production use cases: vLLM behind a drop-in OpenAI Chat Completions API, with LoRA adapters, FP8 quantization, per-use-case isolation, KEDA autoscaling, and end-to-end observability
- Migrated production workloads off external LLM vendors onto the platform, saving over \$7M annually, including an 87% reduction (~\$5.8M/year) on a 60M request/day generation workload and a 93% reduction (~\$1.34M/year) on a fine-tuned ranking model
- Delivered inline query embeddings on the live job-search path at ~20ms p95, a latency budget third-party embedding APIs (~500ms p90) could not meet
- Productionized a fine-tuned SLM reranker at flat online latency, making a 44% NDCG@10 gain and 63% fewer bad results shippable
- Spearheaded cost-saving initiatives delivering over \$3.3M in annual savings, including sunsetting a legacy email alerts product (\$1.7M/year), cutting the team's AWS account costs (\$1.4M/year), and removing unused data builders (\$180K/year)
- Traced a GPU cost-attribution bug misreporting spend across every GPU deployment in the company, driving the cross-team investigation and proposing the fix
- Led the bidding model migration onto the shared model-serving platform as single point of contact, using shadow tests for safe rollout and consolidating bidding onto a global cluster that simplified deployments and improved CPU utilization
- Owned and resolved an 11-month availability failure in a business-critical data pipeline, coordinating 3 engineers and an intern across 5 partner teams, migrating 3 of its indexes to AWS, and restoring the model refresh cadence to daily
- Stepped up as team lead, mentored and onboarded numerous engineers and an intern, co-created a new pair programming interview, and fostered a culture of knowledge sharing and growth
- Set the team's AI-assisted engineering practice, running demos and knowledge-sharing sessions on a spec-driven, agent-assisted workflow that other engineers adopted as their own

Java Spring AWS Kotlin Javascript React Python Docker Scala Kubernetes vLLM SLM LoRA KEDA GPU

Senior Software Engineer

2020 - 2021

Austin, TX | Remote

- Engineered a serverless application empowering patients and clinics to manage communication preferences
- Enhanced application performance by implementing query optimization, pagination, and data caching strategies
- Spearheaded the migration to a newer, more secure version of Spring Boot, mitigating security vulnerabilities

Java Spring Hibernate JPA MySQL AWS Lambda DynamoDB TypeScript

Lead Software Engineer

2017 - 2020

SF, CA - Denver, CO

- Led a team of four developers, providing technical guidance and overseeing project execution to success
- Expanded market reach by integrating with external hotel, car, insurance, and flight systems in multiple countries
- Architected and implemented a new system design to accelerate the integration of third-party APIs

Java Spring Hibernate PostgreSQL Selenium Ember.js FTL

Software Engineer

2011 - 2017

Monterrey, NL | Remote

- Participated in multiple technology migrations including SVN to Git, Maven to Gradle, and Java 7 to Java 8
- Optimized the development feedback loop by parallelizing CI/CD pipelines using Jenkins
- Contributed to the development and enhancement of a travel platform system

Java Spring Hibernate PostgreSQL Selenium Ember.js FTL

Software Engineer

2009 - 2011

Monterrey, NL | Remote

- Established a new business line by building an e-invoice system that integrated with a CMS and regulatory entities
- Implemented a proper branching strategy that consolidated different client implementations into one platform
- Contributed to the development and maintenance of a Content Management System (CMS)

Java Spring

Projects

 den MULTI-HOST NIXOS CONFIGURATION - Declarative config for a dev laptop, a family machine and a NAS: Home Manager, encrypted secrets, CI-checked builds - Assembles every host from auto-discovered flake-parts modules using the dendritic pattern NixNixOSFlakesflake-partsHome Managersops	2025 - Present Remote	 VPN Containers OPEN SOURCE CONTAINER IMAGES - Containers that make routing traffic through the NordVPN, NordLynx and ProtonVPN network easy and secure 18M+ downloads 1K+ stars 250+ forks ContainerizationDockerPodmanIptablesNetworkingBashLinuxWireguardOpenVPNs6	2018 - 2025 Remote
 JUPR UNIVERSAL RATING PLATFORM Implemented a Whatsapp bot that scans messages in group chats, parses matches and exports them to DUPR RustRegexPostgreSQLSeaORMWhatsappPickleballDUPR	2024 Remote	 Rocks 1ST PLACE WINNER AT UNEARTHED DENVER 2017 - Designed a system to filter out low-quality minerals using machine learning and a tomato-sorting-like machine - Trained a convolutional neural network to classify minerals HackathonPythonAndroidGoogle Cloud MLTensorFlowInceptionV3MLCNN	2017 Denver, CO
 Bitcoin Payment Connector PEOPLE'S CHOICE AWARD WINNER - Enabled a travel platform to accept Bitcoin payments - Integrated with Coinbase and TestNet to process transactions HackathonJavaSpringBitcoinBlockchainCoinbaseCryptocurrencyTestNet	2015 SF, CA	 Boardless Games 2ND PLACE WINNER AT NEARSOFT - A multiplayer platform to play board games on phones - Supported games: Love Letter and The Resistance HackathonBoardgamesCoffeeScriptCSSHTMLJadeJavaScriptMeteorMongoDBNode.jsPug	2013 Hermosillo, SO

Open Source Contributions

 genai-bench BENCHMARKING TOOL FOR LLM SERVING Fix SSE parser to handle optional/variable whitespace after "data:"	2026 Remote	 NixOS DECLARATIVE LINUX DISTRIBUTION Update cloudflare-wrap enrollment Add ivsc-firmware to ipu6 option Add a1_prod symbolic links to fix ipu6 webcams	2024 Remote
 Distrobox ANY LINUX DISTRIBUTION INSIDE YOUR TERMINAL Upgrade only running instances Ignore net files when unsharing nets 7 merged PRs 13K+ stars	2023 Remote	 Gluetun VPN CLIENT CONTAINER FOR MULTIPLE PROVIDERS Add NordVPN API endpoint and WireGuard support Add generics to some methods 15K+ stars	2023 Remote
 Datafaker LIBRARY TO GENERATE FAKE DATA Implement a lazy cache on all locales, speeding up to 50% tests using multiple languages	2022 Remote	 TestNG JAVA TESTING FRAMEWORK Fix issue #92 - @BeforeTest method in a super class will be called multiple time when alwayRun = true	2011 Remote

Certificates

2013	Oracle Certified Professional, Java EE 5 Web Component Developer , Oracle
2011	Oracle Certified Professional, Java SE 6 Programmer , Oracle
2019	Certified SAlFe® 4 Agile Software Engineer , SAlFe by Scaled Agile, Inc

Education

 Bachelor of Science - Computer Science FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS, UANL Courses: Software Engineering Design Patterns Database Systems	2005 - 2009 Nuevo León, MX
--	-------------------------------

Skills

Languages	English (Full Professional) Spanish (Native)
Personal Interests	Boardgames Pickleball Self-hosting Linux
AI / Inference	vLLM SLM LoRA FP8 Quantization KEDA GPU Serving
AI Tooling	Claude Code Gemini Agentic Coding Workflows Spec-Driven Development MCP
Tech Stacks	Java Spring Python Kubernetes AWS Terraform Datadog SQL Bash Kotlin Rust TypeScript