Marcus B.
Senior AI/ML Platform Engineer

SUMMARY

- Senior AI/ML Engineer with 15 years in AI/ML and 24 years in systems engineering.
- Worked at Google as a Staff Site Reliability Engineer and at IBM as an AI Research Development Lead.
- Experienced in Python, PyTorch, TensorFlow, FastAPI, SQL, and distributed data pipelines for ML development.
- Owned the ML lifecycle, including model training, inference, evaluation, experiment tracking, drift monitoring, and versioned rollouts.
- Built RAG and retrieval systems using embeddings, hybrid search, Qdrant, Pinecone, and Milvus, processing hundreds of millions of embeddings.
- Holds an MEng in Artificial Intelligence and Computer Engineering from MIT School of Engineering.

TECHNICAL SKILLS

Main Technical Skills	Python (14 yr.), FastAPI, RAG, Kubernetes (17 yr.), Go (8 yr.)
AI Tools & Assistants	Gemini
Programming Languages	C#, C++, Go (8 yr.), Java (5 yr.), JavaScript (3 yr.), Kotlin (3 yr.), Lua (3 yr.), Python (14 yr.), Ruby (13 yr.), Rust (3 yr.), Scala, TypeScript (4 yr.)
AI & Machine Learning	Agent2Agent Protocol, AIOps, Apache Mahout, AWS SageMaker (Amazon SageMaker), Azure AI Foundry, Azure OpenAI, BM25 (3 yr.), ColBERT-style reranking (3 yr.), CrewAI, Cross-encoder scoring (3 yr.), Gemma, GenAI, GPU-aware scheduling, GraphRAG, Hugging Face, INT4 quantization, ISO/IEC 42001, JAX, Keras, KServe, Kubeflow, LangChain, LangGraph, LlamaIndex, LLM gateway, LLM gateways, Machine Learning, ML evaluation harnesses, Mlflow, NVIDIA Triton Inference Server, OpenAI, Pinecone, PyTorch, RAG, Ray, TensorFlow, TensorRT-LLM, Triton, Vertex AI, vLLM
Scala Libraries and Tools	Apache Flink (4 yr.)
Java Frameworks	Apache Spark (4 yr.), Spring Boot
Scala Frameworks	Apache Spark (4 yr.)
Python Frameworks	Django, FastAPI
Python Libraries and Tools	JAX, Keras, Matplotlib, Pandas, PySpark, PyTorch, Ray, TensorFlow
Java Libraries and Tools	JVM, Lucene
	Kotlin (3 yr.)

Android Libraries and Tools	
JavaScript Frameworks	Node.js (3 yr.), React
Data Analysis and Visualization Technologies	Apache Airflow, Apache Hive, Apache Mahout, Apache Nifi, Apache Spark (4 yr.), ETL, Jupyter Notebook, Kafka Connect, Kibana, MapReduce, Pandas
Security	Cosign, CycloneDX, Hashicorp Vault, Identity and Access Management, ISO/IEC 42001, Policy as code, RBAC, Semgrep, Sigstore, SLSA provenance, SOC 2, SPDX
Databases & Management Systems / ORM	Apache Hive, Apache Spark (4 yr.), Bigtable, Cassandra, Elasticsearch, Google Cloud Bigtable, Google Cloud Spanner, MySQL (8 yr.), Neo4j, NoSQL, Oracle Database, PostgreSQL (10 yr.), RDBMS, Redis (8 yr.), SQL (10 yr.)
Cloud Platforms, Services & Computing	AWS (3 yr.), Azure (4 yr.), GCP (6 yr.), OCI (4 yr.)
Amazon Web Services	AWS Auto Scaling (3 yr.), AWS Bedrock, AWS EventBridge, AWS SageMaker (Amazon SageMaker)
Azure Cloud Services	Azure AI Foundry, Azure Event Hubs, Azure Machine Learning, Azure OpenAI, Azure Service Bus
Google Cloud Platform	Dataflow, Google BigQuery, Google Cloud Bigtable, Google Cloud Dataflow, Google Cloud Pub/Sub, Google Cloud Spanner
Industry Domain Experience	HIPAA
SDK / API and Integrations	Agent2Agent Protocol, API, FastAPI, GraphQL, LLM gateway, LLM gateways, OpenAPI, OpenTelemetry (3 yr.)
Methodologies, Paradigms and Patterns	Agile, GDPR, IaC, REST, Scrum
Deployment, CI/CD & Administration	AIOps, Ansible (8 yr.), ArgoCD (3 yr.), Chef (8 yr.), CI/CD (4 yr.), DORA Metrics, Flux, GitOps (5 yr.), Gradle, Helm (4 yr.), Jenkins (5 yr.), loki (3 yr.), Policy as code, Pulumi, Puppet (8 yr.), SLSA provenance, SonarQube
Tools for Big Data	Apache Arrow (3 yr.), Apache Avro, Apache Iceberg, Google Cloud Dataflow
Web/App Servers, Middleware	Apache Kafka (5 yr.), Nginx (8 yr.), Varnish (9 yr.)
QA, Test Automation, Security	Apache Maven, Bazel
Platforms	Apache Mesos, Red Hat OpenShift Container Platform
Collaboration, Task & Issue Tracking	Atlassian Confluence, Jira
Scripting and Command Line Interfaces	Bash, Perl, Shell Scripts (11 yr.)

Mail / Network Protocols / Data transfer	BGP (8 yr.), DNS (8 yr.), GRPC, TCP/IP (8 yr.)
Logging and Monitoring	Datadog, Grafana, Grafana Loki (3 yr.), Logstash, perf, Prometheus (3 yr.)
Virtualization, Containers and Orchestration	Docker (4 yr.), GPU-aware scheduling, Kubernetes (17 yr.), KVM (for Kernel-based Virtual Machine) (8 yr.), Linux cgroups, Linux namespaces, LXC (8 yr.), Packer, Terraform (5 yr.), Vagrant, VmWare (8 yr.), XEN (8 yr.)
Version Control	Git, GitHub, Github Actions
Message/Queue/Task Brokers	Kafka (5 yr.), RabbitMQ
Operating Systems	Linux, Linux cgroups, Linux Kernel, Linux namespaces, Unix
Third Party Tools / IDEs / SDK / Services	Trivy
Other Technical Skills	Amazon MSK, Apache Parquet (3 yr.), AutoGen, Bicep, eBPF, Flink (4 yr.), Google Kubernetes Engine, JFrog Artifactory, LGPD, LoRa, Milvus (3 yr.), OpenStack (8 yr.), Oracle Cloud Infrastructure (4 yr.), PGVector (3 yr.), Qdrant, Semantic Kernel

WORK EXPERIENCE

CVS Health, Senior AI Cloud Security Engineer - Colossus Tech Group

Aug 2026 - Present

Through Colossus Technologies Group, designs enterprise-scale cloud and AI security solutions for a regulated healthcare estate, including guardrails and governance for GenAI workloads and identity and policy controls across multi-cloud Kubernetes.

The Apache Software Foundation, Senior Technical Lead

Jun 2026 - Present

- Leads technical direction for six contributor teams across Kafka, Spark, Iceberg, NiFi, Mahout and Mesos.
- Migrates JVM projects to current Java LTS releases by replacing deprecated and removed APIs, updating Maven and Gradle toolchains, and validating builds across Java versions.
- Enforces backward compatibility and stable public APIs through clear deprecation policies and semantic versioning so downstream users can upgrade without breakage.
- Reviews vendor pull requests for project neutrality and community-driven rather than single-vendor roadmaps. Standardizes release engineering through reproducible builds on GitHub Actions CI, signed artifacts and compatibility gates that reduce manual release work.
- Runs supply-chain security, including SBOMs, dependency and CVE triage tracked in ASF Jira and GitHub, and coordinated disclosure for projects used in thousands of deployments.

- Raises test coverage and reduces flaky tests to accelerate contributor reviews and merges. Mentors committers and improves review and roadmap practices to maintain long-lived projects.
- Directed an OpenSSF-aligned release and software-supply-chain program spanning Java, Kotlin, Scala, Rust, Python and Go projects. Used Maven, Gradle, Bazel, GitHub Actions, SLSA provenance, CycloneDX and SPDX SBOMs, Sigstore and Cosign signing, SonarQube, Semgrep, Trivy, dependency policies, reproducible builds, fuzzing, chaos tests, OpenTelemetry telemetry, and Kubernetes validation across AWS, Azure and GCP.

SM4RT WORKS, Staff MLOps Engineer/ML Infrastructure Architect

May 2026 - Jun 2026

- Shipped customer-facing GenAI features from prototype to production, including spend and document intelligence, automated actions, and agentic workflows using FastAPI, Python, LangChain, LangGraph, OpenAI and Hugging Face, in partnership with product engineers and data specialists.
- Built multi-step, tool-using LangGraph agents with function and tool calling, retrieval, stateful memory, and human approval for high-stakes actions, integrated into a user-facing product running at 99.9% reliability.
- Owned evaluation, monitoring and operationalization of LLM features. Defined the team's LLMOps practice through offline and online evaluation harnesses, production quality, drift, latency and cost tracking, and versioned prompt and model changes behind canary rollouts.
- Engineered RAG and data-retrieval pipelines using vector and hybrid search with Ray, Qdrant, Pinecone and Milvus, fed by SQL, BigQuery and Airflow. Owned data quality, schema and retrieval architecture end to end.
- Partnered with the platform and GenAI-core function on shared LLM gateways, model routing, guardrails, evaluation and observability. Explained AI trade-offs to Product, Design and business stakeholders.
- Architected and built production AI/ML systems, distributed backend services and microservices on Kubernetes for mission-critical, regulated environments, focusing on reliability, model accuracy, RBAC, audit trails and DLP.
- Owned the ML lifecycle as a platform, including model-training workflows, real-time inference through KServe, vLLM and Triton, continuous drift and quality monitoring, troubleshooting, model-management tooling, and versioned, auditable rollouts.
- Defined and enforced latency, throughput and resource-utilization SLAs using GPU-aware scheduling, autoscaling, error budgets, and FinOps cost and performance tuning.
- Engineered model deployment using Docker, Kubernetes, custom Go operators, Helm, Terraform and GitOps, with GitHub Actions CI/CD and versioned model and container artifacts in JFrog Artifactory. Built across data and software layers in Python, Go, TypeScript and Node.
- Led cross-functional initiatives, planning scope, timelines and dependencies in Jira and documenting them in Confluence. Aligned Data, Infrastructure and ML Engineering, serving as the ML-infrastructure subject-matter expert for internal and external stakeholders.
- Built SQL, NoSQL, Spark, PySpark, Kafka, Airflow and dbt pipelines feeding the ML platform, and delivered LLM gateways, RAG and vector-search features. Mentored engineers and promoted knowledge-sharing and continuous improvement.
- Built a regulated multi-cloud AI control plane spanning Amazon Bedrock and SageMaker, Azure AI Foundry, Azure OpenAI and Azure Machine Learning, and Vertex AI and Gemini. Included MCP and A2A agents, GraphRAG, PyTorch, JAX and TensorFlow training, MLflow and Kubeflow governance, vLLM, Triton and TensorRT-LLM inference, NVIDIA GPU scheduling, Kafka, Flink and Spark lakehouse pipelines,

model-risk controls, EU AI Act evidence, zero-trust tenancy, SLOs, FinOps, disaster recovery, and human approval for high-impact decisions.

QAT Global

Jan 2026 - Apr 2026

Staff Software Engineer | Infrastructure Specialist

Jan 2026 - Apr 2026

- Architected a dual-platform agentic GenAI system using LangGraph, Python, FastAPI, Kafka and Flink for autonomous agent provisioning across isolated tenants, reducing cross-platform integration effort by 60%.
- Shipped a production LLM gateway with model routing, token metering, rate limits, retries, circuit breakers and provider failover. Deployed self-hosted inference using Gemma-class models, LoRA, vLLM-style batching, KV cache and INT4, achieving 3.2x GPU throughput at less than 2% accuracy loss.
- Engineered LGPD-first architecture aligned with GDPR, NIST AI RMF, ISO 42001 and SOC 2, with reusable Terraform, Helm and Kubernetes templates for multi-tenant AI services across AWS, GCP and Azure. Coordinated client delivery in Jira and Confluence.
- Delivered a sovereign agent platform using Python, Go, TypeScript, Java, C#, FastAPI, React, gRPC, GraphQL, OpenAPI, Kong, LangChain, LangGraph, Semantic Kernel, AutoGen, CrewAI, LlamaIndex, vector databases and knowledge graphs. The source also lists a malformed .NET technology reference.
- The sovereign platform included Azure Service Bus and Event Hubs, AWS EventBridge and MSK, GCP Pub/Sub and Dataflow, OpenShift, Kubernetes, Crossplane, Terraform, Pulumi, Bicep, Argo CD and Flux, policy as code, model evaluations, prompt-injection defenses, audit trails, and active-active regional failover.

Staff Software Engineer | Infrastructure Specialist - PowerOmni

Built PowerOmni, a one-click agent-provisioning service using FastAPI, Redis, Qdrant and tenant-aware templates, reducing agent deployment time from hours to under 90 seconds.

Staff Software Engineer | Infrastructure Specialist - Foundry

Built Foundry, a pipeline-materialization service using NiFi, Kafka, Flink, GitHub Actions CI and ArgoCD GitOps to create, version, roll back and tear down AI and data pipelines in real time across AWS and GCP.

AI4Design, Tech Lead DevOps/SRE & Innovation | Staff AI Engineer

Jan 2023 - Dec 2025

- Architected hybrid RAG using BM25, Lucene, Qdrant, pgvector, Milvus, ColBERT-style reranking and cross-encoder scoring, with real-time indexing through Kafka, Flink, Debezium CDC, Spark, Arrow and Parquet. Reduced content freshness from hours to minutes across hundreds of millions of embeddings.
- Designed a GitOps-driven multi-tenant AI platform using Kubernetes, ArgoCD, Terraform, Helm, Go operators and Python automation across AWS, GCP, Azure and OCI. Reduced deployment lead time by 40% and eliminated environment drift.

- Built a polyglot platform control plane in Python, TypeScript, Node.js, Go, Rust, Java, Kotlin, C, Ruby, Lua and shell, spanning APIs, CLIs, Kubernetes controllers and migration tooling. Standardized 40+ services and reduced integration effort by 35%.
- Established evaluation and LLMOps practices, including evaluation harnesses, golden datasets, drift, quality and cost monitoring, and safe prompt and model-version management through CI/CD and GitOps. Reduced model iteration from weeks to days.
- Integrated AI-driven autoscaling, anomaly detection and event-driven automated remediation using Prometheus, Loki, OpenTelemetry and Kubernetes controllers, reducing MTTR by 45% and maintenance overhead by 35%.
- Ran SRE practices covering SLIs, SLOs, error budgets, incident reviews, runbooks, capacity planning and FinOps dashboards, keeping customer-facing GenAI features reliable and production costs visible.

Kubernetes, ArgoCD, RAG, MLOps, Site Reliability Engineering

Globo, Chief Architect of Cloud Infrastructure Projects

Jan 2022 - Dec 2022

- Created a Go traffic router for 280 CDN points of presence, including health checks, geolocation, weighted routing and consistent hashing, improving regional failover and eliminating manual traffic operations.
- Delivered the first incident-free, two-round national election coverage since 2010. Neutralized a multi-thousand-node botnet with a Varnish tarpit holding connections in about 2 MB of memory, and added linear-extrapolation circuit breakers and exponential backoff across multiple origins to avoid HTTP 429 responses from government APIs.
- Automated cloud services and DBaaS provisioning across GCP, OCI and private cloud using infrastructure as code and policy-driven templates, reducing operational cost by 50% while improving security and capacity planning.
- Led cloud infrastructure modernization using Kubernetes, Terraform, GitOps, Jenkins, managed databases and reusable platform modules, delivering the full project while improving repeatability, auditability and release confidence.
- Built a platform metadata and knowledge graph layer with Neo4j and Kafka Connect, improving service ownership visibility, dependency mapping and contextual retrieval for architecture and operations decisions. Optimized CDN and web performance through Varnish caching and capacity tuning.
- Partnered with engineering, security and operations leaders to translate platform risks into roadmaps, migration plans and measurable reliability objectives.

RDBMS, Hive, RESTful WebServices, API Development, Redis, Bash, Docker, Jenkins, SQL, Ansible, Cloud Computing IaaS, AIOps, Microservices, GCP, CI/CD, Kubernetes, Varnish

Platform - Spotify, Staff Software Engineer

Apr 2020 - Dec 2021

Core engineer on Spotify's internal developer platform as Backstage became open source. Expanded its software catalog, scaffolder, TechDocs docs-as-code and custom plugins, consolidating thousands of components, services and data pipelines under clear ownership metadata. Developed custom plugins in TypeScript and React.

Built golden-path scaffolder templates that reduced new-service creation from days to under an hour, with repository bootstrap, CI/CD wiring, Kubernetes manifests, artifact publishing, alerting and production-readiness checks included.

Drove IaC and GitOps adoption across squads through reusable Terraform modules for GCP and GKE, declarative deployments with drift detection, and policy-as-code checks in CI, reducing environment provisioning from weeks to hours.

Worked on an event-oriented delivery platform using GCP Pub/Sub, a schema registry and Avro, ingesting billions of events per day into data and ML pipelines with dead-letter queues, backpressure handling and end-to-end delivery SLOs.

Instrumented DevEx and DORA metrics, including lead time, deployment frequency and MTTR, with adoption dashboards used to prioritize the platform roadmap. Mentored engineers on paved-road patterns and inner-source contributions to Backstage plugins.

Tech Lead (L6 - SRE / Staff SWE) - Google, Staff Site Reliability Engineer

Jul 2018 - Mar 2020

- Refactored and extended Kubernetes/Borg-style scheduling components in Go using priority, preemption, bin-packing, quotas and topology-aware placement, improving multi-tenant resource allocation in edge cases involving clusters above 200,000 nodes and 100M+ resources.
- Served as guest architect and developer for Gecko and GCP projects, drawing on Borg experience and advocating Kubernetes.
- Applied Linux internals, including cgroups, namespaces, scheduling, memory isolation and container runtime behavior, to improve workload stability and hardware isolation across production clusters.
- Owned SLI, SLO and error-budget frameworks for critical production services. Combined incident reviews, postmortems, capacity planning and reliability automation to improve scalability and operational predictability by roughly 30%.
- Standardized CI/CD and production releases using Jenkins, Docker, Kubernetes, automated validation and staged rollouts, reducing release risk and tightening feedback loops between product engineering and SRE.
- Optimized high-throughput data platforms through Cloud Spanner and Bigtable tuning, schema and access-pattern analysis, and production load testing, improving latency and throughput for globally distributed workloads.
- Designed systems-level performance improvements for low-latency inference and user-facing services by balancing memory management, scheduling, request batching, and accuracy and latency trade-offs.
- Led cross-functional reliability work with software engineers, product teams and infrastructure specialists, translating ambiguous production risks into engineering plans and measurable SLO outcomes using error budgets. Adopted Blue Book practices.

RDBMS, Redis, Big Data Analytics, RESTful architecture, Bash, Docker, SQL, Cloud Computing IaaS, AIOps, Kafka, GCP, CI/CD, Go, Kubernetes, Linux, AWS, Python

Learning Models - IBM, AI Research Development Lead

Aug 2017 - Jul 2018

- Led ML systems research and prototype delivery using Hadoop, Spark, Python, Jupyter and distributed compute pipelines, accelerating training and experimentation cycles by 60%.
- Designed reproducible model-development workflows with notebooks, dataset versioning, testable feature pipelines and CI validation, reducing regression defects by 50% and improving researcher-to-production handoff.
- Built distributed data-preparation and model-evaluation pipelines using Spark, Python and GPU-aware orchestration, improving repeatability across batch training and offline

evaluation. Introduced automated testing and quality gates, increasing test coverage by 50% and reducing model and pipeline regressions by 40%.
- Created visualization and analysis tools using Python, Pandas-style workflows and Matplotlib-compatible reporting, reducing exploratory analysis time by 30% for research and engineering teams.
- Partnered with researchers, engineers and product stakeholders to turn experimental learning-model ideas into measurable prototypes, architecture documents and production-readiness plans and recommendations.

RDBMS, JavaScript, MapReduce, Natural Language Processing (NLP), Hive, Neo4j, API Development, Predictive Modeling, Big Data Analytics, Data Science, RESTful architecture, Bash, Docker, Jenkins, SQL, Cloud Computing IaaS, AIOps, Kafka, TensorFlow, Machine Learning, Artificial Neural Networks, CI/CD, Python

Tradeshift, Senior DevOps Engineer

Nov 2015 - Jul 2017

- Improved the software-development process through Jenkins-based CI/CD and microservices architecture, increasing delivery speed by 20%.
- Built automation tooling in Go.
- Integrated automated regression testing into the Jenkins pipeline, reducing testing time by 60% and increasing release frequency by 20%.
- Led a cross-functional team implementing caching and CDN integration, reducing server load by 70% and increasing capacity to 100,000 concurrent users.
- Harmonized SQL and NoSQL data models for a hybrid e-commerce platform, improving transactional efficiency and scalability.
- Developed a multi-language application suite using Java for backend logic, Python for data processing and Ruby for scripting.
- Containerized the suite with Docker for consistent runtimes and deployment across platforms.
- Hosted and tuned applications on Linux servers for uptime and performance.
- Secured sensitive application secrets and credentials with Vault.
- Collaborated on replicable development environments with Vagrant and custom server images with Packer, streamlining the development-to-deployment pipeline.
- Integrated Cassandra for high-write scenarios with PostgreSQL's ACID properties to support analytics and transactions.
- Implemented the ELK stack, including Elasticsearch, Logstash and Kibana, to improve search and visualization of operational data.
- Combined metrics, traces and logs in Datadog to provide a holistic view of system health and performance benchmarks.

RDBMS, RabbitMQ, JavaScript, Neo4j, Spring Boot, Bash, Docker, Jenkins, SQL, Ansible, Elasticsearch, AIOps, CI/CD, Go, Kubernetes, Terraform, AWS

Hotel Urbano, Staff Engineer Development Operations

Oct 2014 - May 2015

- Improved infrastructure and DevOps practices with Jenkins through automated provisioning and code deployment, reducing deployment time by 30%.
- Automated infrastructure with Puppet across a distributed network of 200+ servers, reducing provisioning time by 75%.
- Led development and implementation of deployment automation using Jenkins and Ansible, reducing manual deployment errors by 75% and saving 20+ developer hours per week.

RDBMS, Redis, Bash, Docker, Jenkins, SQL, Ansible, Elasticsearch, AIOps, GCP

Globo.com, Staff Engineer / Operating Systems Specialist

Jan 2007 - Sep 2014

Led a distributed infrastructure-automation platform for datacenter virtualization and IaaS, increasing server utilization by 30% and reducing energy consumption by 50%. Built pre-Kubernetes self-service orchestration. Operated and tuned Unix/Linux, SAN storage and Varnish caching for high-throughput, mission-critical media workloads. Built shell, Ruby and Python automation and mentored engineers on operating systems, Unix internals and production diagnostics. Designed an early cloud and SRE platform for national-scale media using KVM, Xen, VMware, OpenStack, LXC containers, Puppet, Chef, Ansible, SAN, NAS and object storage, BGP, DNS and TCP/IP load balancing, Nginx, Varnish, CDN, MySQL, PostgreSQL, Redis, Hadoop, message queues, monitoring and log aggregation, capacity models, high availability and geographically distributed disaster recovery.

Globo

7 years 9 months

Staff Engineer

Sep 2012 - Sep 2014

Led design and development of a distributed system to automate and manage IT infrastructure, enabling full datacenter virtualization for an IaaS project. Optimized resource allocation, increasing server utilization by 30% and reducing energy consumption by 50%.

RDBMS, API Development, Bash, Docker, SQL, AIOps, CI/CD, Linux, Cloud Computing, Virtualization

Operating Systems Specialist

Jan 2007 - Sep 2012

Troubleshot and developed solutions for Unix services, administered SAN storage, and implemented and administered Varnish for large HTTP workloads, supporting web developers.

RDBMS, API Development, Bash, SQL, AIOps

From Residuals to Returns - Quantitative Paper on Global capital Market

Jan 2026 - Feb 2026

Quantitative research paper on global capital markets, published on SSRN as From Residuals to Returns: A Multi-Variable, Regime-Aware Framework for Probabilistic Market Inference on July 27, 2026.

Presents a framework for probabilistic market inference based on residuals, or idiosyncratic price movements remaining after dynamic removal of macroeconomic factors through robust Kalman filtering, rather than raw equity returns. Integrates state-space models with Student-t robust extensions into a unified pipeline.

Platform Entities GenAI

Jan 2026 - Feb 2026

AI platform strategy material concerning iFood describes Domain-Driven Design, hyper-personalized recommendations, autonomous dispatch optimization, real-time fraud detection, restaurant tools intended to reduce food waste and improve margins, and conversational support AI. The material cites product context of 80M+ monthly orders and 330K+ restaurant partners, and a three-year roadmap targeting 356% ROI. These are strategy and product-context statements; individual implementation responsibility and achieved results are not specified.

Athenas Active Incident Management.

Jan 2024 - Sep 2025

Athenas is an automated incident-response system using collaborative software agents for proactive, evidence-based remediation. Its documentation also presents Batteries Included, a source-available platform integrating monitoring, security and serverless tools for Kubernetes management, and emphasizes well-defined interfaces for scalable, robust operations. Individual implementation responsibility is not specified.

DAISY - Horizontal AI

Apr 2017 - Oct 2017

Conceptual description of an AI system with three separate neural networks of 50 layers each. The proposed functions are feature extraction, sequence generation or processing, and classification or regression. The architecture is presented as hypothetical and dependent on the problem, not as a confirmed implementation.

Neural Network 1 - Feature Extraction

The first network could transform raw inputs into an intermediate representation. Suggested options include a CNN for images, identifying lines, shapes and textures, or an RNN or Transformer for sequences such as text or time series.

Neural Network 2 - Sequence Generation/Processing

The second network could process the feature-rich representation and generate a sequence, for example using an RNN or Transformer to produce target-language words in machine translation.

Neural Network 3 - Classification/Regression

The third network could be feed-forward and use the second network's output for final classification or regression predictions, such as identifying an image object or determining text sentiment.

Kubernetes

Apr 2013 - Jun 2015

The project description explains Kubernetes as a distributed container-orchestration system that deploys, scales and manages containerized applications using a declarative model and self-healing control plane. It abstracts compute, networking and storage across a cluster and maintains the desired workload state. No individual contribution is specified.

Varnish Group Meeting - Presentation and Workshop

Oct 2012 - Nov 2012

Varnish presentation and workshop involving a hackathon and technical discussion.

Google Search - Optimizations fine tuning and creativity

Optimized and tuned a search stack covering crawlers, massive indexes, queues, SEO, serving, search and reranking.

Helping CERN

Worked on a large Python codebase at CERN described as containing over a million bugs, identifying, diagnosing and resolving complex software issues in a large-scale, mission-critical environment. The bug count describes the codebase, not the number personally resolved.

k8s-microlens

Kubernetes insights and resource-visualization tool, also listed as K8s MicroLens. Written in Go using Pomodoro, it presents cluster resources and their relationships hierarchically.

Nexus - Ads platform

Advertising-platform project described as revisiting the advertising market.

Omni Agentic IDP Platform

Video overview of Omni-Agentic Platform IDP, an internal developer platform presented as a macro-kernel operating environment for enterprise AI execution. It addresses AI systems as distributed software requiring multi-cloud infrastructure, security and memory management. Individual delivery responsibility is not specified.

Varnish Infrastructure

Maintained and administered reverse-caching infrastructure based on Varnish.

Multi-Stage Neural Networks with Learnable Intermediate Bridges: From a Theoretical Framework to Production Retrieval-Augmented Generation

Paper published on SSRN on August 15, 2026. Formalizes compositional systems of three neural networks connected by learnable intermediate bridges: trainable interface modules governed by explicit contracts on shape, semantics, distributional stability, information content and computational budget. Extends the 2018 framework to production retrieval-augmented generation.

EDUCATION

- MIT School of Engineering
  Master of Engineering in Artificial Intelligence and Computer Engineering.
  2023–2025
- Federal University of Rio de Janeiro (UFRJ)
  Postgraduate degree in Psychology.
  2021–2023

- Federal University of Rio de Janeiro (UFRJ)
  BSc in Computer Science.
  2005–2011
- Federal University of Rio de Janeiro (UFRJ)
  BSc in Applied Mathematics.
  2007–2011
- Federal University of Rio de Janeiro
  Postgraduate Degree in Psychology; Psychology
  2021 – 2023

CERTIFICATION

- **Certified Information Systems Security Professional (CISSP)**
  Apr 2020 - Apr 2030
- **CKS (Kubernetes Security)**
  Cloud Native Computing Foundation
- **Red Hat Certified Engineer (RHCE)**
- **Red Hat Certified System Administrator (RHCSA)**
- **Red Hat Certified JBoss Administrator (RHCJA)**
- **Cisco Certified Network Associate Cloud (CCNA-Cloud)**
  Jan 2021 - Nov 2029
- **CCNA Routing and Switching**
  Cisco
- **Cisco Certified Network Professional (CCNP)**
- **Offensive Security Certified Professional (OSCP)**
  Offensive Security
- **PREC-CP**
  US Army Corps of Engineers
- **C|OASP, OSAI, GIAC GOAA, GASAE, GAIPS, GMLE, ISACA AAISM, AAIA, AAIR, and ISO/IEC 42001 Lead Implementer or Lead Auditor**
  Microsoft / ISO
- **6x Oracle Cloud Infrastructure (OCI) Certified**
  Oracle
- **9x IBM Cloud Certified**
  IBM
- **24x Microsoft Azure Certified**
  Microsoft
- **14x Google Cloud Platform (GCP) Certified**
  Google
- **12x Amazon Web Services (AWS)**
  CertifiedAmazon Web Services