# Kozato Keizo

SENIOR FULL-STACK & AI/ML ENGINEER
+ 628983883704 | keizokozato@gmail.com | github.com/lesofger

## SUMMARY

Senior Full Stack & AI Engineer with 9+ years of experience building and delivering scalable, distributed web applications across modern cloud environments.
Proficient in LangChain/LangGraph, LLms, advanced RAG, React/Next.js, Python/FastAPI, Java Spring Boot, and Azure/AWS Kubernetes with deep expertise in microservices architecture, API engineering, and cloud management for seamless deployment and reliability. Integrated AI-driven services into full-stack ecosystems, ensuring high performance, security and maintainability.

## SKILLS

- Programming Languages:  Python, C#, JAVA, Scala, JavaScript/TypeScript
- Frameworks and tools: .NET, Spring Boot, Spring Cloud, Django/Flask/FastAPI, ReactJS/Next.js, AngularJS, Akka, Hadoop, Zookeeper, MapReduce, Apache Spark, Azure Synapse, ETCD
- Machine Learning Technologies: TensorFlow, PyTorch, Hugging Face Transformers, Scikit-learn, NLP, GenAI, LangChain, LangGraph, LLm, RAG, Vector Database (Pinecone, FAISS), Computer Vision
- Mobile Development: React Native, Flutter
- API & Integration: REST APIs, GraphQL, WebSockets, gRPC, RabbitMQ, Kafka
- Databases: MySQL, PostgreSQL, MongoDB, DynamoDB, Neo4j, Cassandra, Redis
- DevOps & Cloud Tools: Azure, AWS, GCP, Cloudflare, CI/CD tools, Git, Docker, Kubernetes, Terraform, Jenkins, Helm, Prometheus, JMeter, Linux, Splunk

## WORK EXPERIENCE

**Senior AI/ML Engineer, Rapids AI** **Jan 2023 – present**

- Directed the development of comprehensive LoRA fine-tuning pipelines using Azure Foundry, PyTorch, and ONNX, covering data preprocessing, model training/validation, packaging, scalable deployment, and operational monitoring of AI models.
- Architected and led production Agentic MCP in Python to orchestrate tools, handle API calls, perform secure database operations, and support advanced RAG features HyDE, Query Decomposition, Weighted Transformation, Text-to-SQL powering 50k+ daily agent interactions at 200ms P95 latency, 99.99 uptime.
- Played a key role in optimizing search performance, contributing to an 80% increase in speed. This was achieved through a combined effort in fine-tuning Elasticsearch queries and enhancing the search APIs. My involvement focused on optimizing query performance, constructing efficient queries, and contributing to API scalability.
- Conducted research and benchmarking on chunking strategies and embedding techniques to enhance semantic preservation in long-form documents, improving LLM pipeline performance and consistency.
- Investigated in backend services in FastAPI with scalable structures, integrating with Azure SQL Database and MongoDB, Redis-backed multi-level caching, for smooth integration of AI pipelines in production.

- Developed robust, production-grade systems by integrating Python's core libraries—dataclasses, functools, itertools, collections, and logging—ensuring maintainable code, improved readability, predictable performance, and enterprise-level reliability across distributed microservices and ETL pipelines.
- Orchestrated a microservice architecture facilitating distributed APIs and event driven communication through Kafka and internal message queues, decreasing average API latency by 30%.
- Unified AI Studio, an Electron + React.js platform with a .NET backend integrating chat streaming, RAG context injection, Azure services and model orchestration.
- Established Kubernetes platform AKS with Terraform, network policies, cluster auto scaling rules and GitOps-style deployment.
- Delivered full-stack observability Prometheus + OpenTelemetry + Jaeger and testing culture pytest + playwright + chaos, validated APIs with mocks, enforced coverage, and implemented health check endpoints for monitoring.
- Engaged in learning and applying new technologies, including Elasticsearch, to better contribute to our search capabilities, emonstrating a commitment to professional growth and staying abreast of evolving industry trends

**Full Stack Developer, PingWind** **Nov 2021 – Jan 2023**

- Implementing ReactJS and Redux to create user interfaces and proficient in leading frameworks such as React.js to create high-quality, scalable, and reusable components and front-end solutions. ReactJS was used to build specific components for data manipulation and presentation in the company's standard format.
- Used React.js and Redux framework for premade components from NPM and worked with Redux architecture increase website speed, as well as React Flux architecture and Redux form to handle form state in Redux.
- Develop and execute all phases of testing with Jasmine, Karma for unit testing, and Protractor for End2End automated testing on both the client and server sides utilizing the Mocha and Chai frameworks.
- Implement different Design Patterns including MVVM for WPF and Silverlight applications and used J2EE Design Patterns Session Facade, Aggregated entity for the Middle Tier development and developed EJBS Session and Message-Driven Beans in RAD for handling database access and asynchronous messaging.
- Leveraged features of Java 8 such as Default and Static methods in Interfaces, Lambda expressions, StreamAPI, Nashorn JavaScript Engine, IO Enhancements, and the Parallel sort in Arrays.
- Migrated SOAP web service resources to Spring RESTful utilizing the Spring REST API and Spring Boot and created action classes in Struts and developed managers classes using Hibernate, JPA and Strut technologies and implement application-level persistence using Hibernate and Spring.
- AI Chatbot Development Contribution: Contributed to the company's AI chatbot development, working in tandem with the Data team to integrate advanced LLM models, enhancing user interactions and overall functionality.
- Involved in creating and maintaining the architecture for a Restful API utilizing Spring Boot to monitor and manage the application in a production environment, as well as integrating Spring Security to validate users.
- Implement the Data Access Layer with Spring Data and the Hibernate ORM tool and created a web-application for business reporting, system monitoring and troubleshooting with J2EE and Tomcat.

**Backend Engineer II, Rally Health** **Jun 2019 – Nov 2021**

- Refined Proton, a command-line tool facilitating faster service creation and updates; drove a 40% increase in developer velocity and was adopted as the most used CLI tool within 6 months.
- Innovated Archimedes, a dependency tracker, which identified and helped resolve the three biggest causes of dependency-related build failures, thus improving build stability and reliability.
- Enhanced backend systems with async programming, improving concurrency handling and scalability which overall system performance and stability by 3.5x.
- Built advanced React and Angular dashboards in TypeScript to visualize large-scale product and user-flow analytics, delivering smooth performance across 50K+ data points.
- Engineered an ASP.NET backend powering real-time monitoring tools that visualized service performance, error trends, and dependency health, reducing investigation time by 60%.
- Rectified a critical account auto-refill error, collaborating with over 5 external engineers to prevent erroneous card transactions, and saving the company $100,000 per month previously lost to reimbursements.
- Integrated AWS services (S3, Lambda, Aurora) for cloud-native storage and compute scaling while deploying SaltStack for infrastructure automation, leveraging Kubernetes and Docker for seamless container orchestration, and developing robust CI/CD pipelines that reduced deployment time by 40%.

**Full Stack developer,** Infosol Inc **Sep 2017 – May 2019**

- Discussed, analyzed, and strategized product design with Product Managers, Design Team, and Business Managers.
- Created and deployed RESTful API endpoints using NodeJs, ASP.NET, and InfoBurst cached XDC queries on AWS.
- Developed and designed web application visualizations using JavaScript frameworks and libraries such as Angular 5/6/7 and React w/ Redux.
- Transformed SAP BusinessObjects reports from SAP - Web Intelligence into consumable data for API use for multiple client projects.
- Created SAP BusinessObjects dashboards proof of concepts for multiple Business Intelligence clients.
- Crafted dashboards and visualization heavy applications using UI and Dashboard frameworks such as Telerik, KendoUI.
- Formulated multiple POC web applications with integrated OKTA and Auth0 in React and Angular for multiple clients.

## EDUCATION

Bachelor’s Degree of Computer Science
Xidian University | Jun 2013 - May 2017