Mark Yamato

Senior Backend Engineer | AI Infrastructure & Microservices | Golang, AWS, Kubernetes

Tokyo ,Japan | +1 210 339 1770 | mrmarkg101@gmail.com | linkedin.com/in/ mark- yamato/.

Summary

Senior Backend Engineer with 7+ years building high-performance distributed systems and scaling platforms from startup to enterprise. Specialized in Golang microservices, database optimization, and production AI integration—pioneered GPT-3 deployment in 2021 before ChatGPT existed. Delivered critical infrastructure during high-stakes scenarios: sustained system stability through WeWork's 2,400-person layoff and BetterHelp's 300% COVID-19 traffic surge. Combines deep technical execution with business impact, reducing operational costs 40-70% while maintaining 99.9%+ uptime across systems serving millions of users.

Professional Experience

ConnexAI | Miami, FL
Senior Software Engineer
March 2023–Present | Hybrid

- Architected end-to-end multi-LLM gateway orchestrating GPT-4, Claude, and PaLM with proprietary routing algorithm, reducing client LLM expenses by 70% across 2M+ monthly requests
- Constructed production RAG system using Pinecone vector database with semantic chunking achieving <300ms P95 latency across multi-million token enterprise document repositories—among earliest enterprise implementations
- Engineered multi-tenancy security architecture with strict data isolation, enterprise SSO integration, automated PII detection via NER models, and audit logging infrastructure for compliance requirements
- Developed proprietary AI function calling framework integrated with LangChain for complex multi-step reasoning—launched in parallel with OpenAI API release to support custom autonomous agent orchestration
- Established engineering excellence initiatives including code review standards and CI/CD improvements driving daily releases, reducing production incidents by 65% while mentoring 3 engineers from junior to mid-level

Backend Engineer
October 2021–February 2023 | Hybrid

- Architected core conversation orchestration service in Golang with WebSocket server managing 500K+ concurrent conversations monthly at <200ms P95 latency with graceful degradation under load
- Integrated OpenAI GPT-3 API in November 2021—one year before ChatGPT launch—building prompt engineering frameworks and context management systems when industry best practices didn't exist
- Built infrastructure including API Gateway with JWT authentication, Redis Pub/Sub for multi-instance routing, DynamoDB state management, and circuit breaker patterns for API failures with fallback responses
- Optimized semantic caching using vector similarity and conversation context windowing reducing OpenAI API costs by 60% across 15 enterprise clients supporting millions of messages monthly

BetterHelp | Deerfield Beach, FL
Backend Engineer
November 2020-August 2021 | On-Site

- Owned complete project lifecycle architecting and deploying Golang microservice replacing legacy matcher, supporting 10X therapist network growth from 3K to 30K profiles with sub-50ms search performance
- Developed weighted scoring system evaluating 14+ compatibility factors with Elasticsearch integration executing complex searches across 30K therapist profiles, improving compatibility scores by 42% via post-session surveys
- Delivered 93% latency reduction for total user matching flow via query optimization and parallel processing; system handled 15K+ matches daily at 99.99% uptime during peak demand periods

- Integrated ML ranking model via REST API with circuit breaker fallback; deployed A/B testing framework showing 23% improvement in 90-day engagement and 28% reduction in first-session cancellations

Full Stack Engineer
March 2020–October 2020 | On-Site

- Diagnosed critical bottlenecks during March 2020 pandemic surge, deploying emergency performance fixes sustaining system stability across unprecedented 100K to 400K concurrent user growth in 3 months
- Rebuilt therapist search UI with React Hooks and virtual scrolling, added Redis caching, replaced synchronous flows with AWS SQS async queuing cutting processing time from 4.2 minutes to 45 seconds
- Sustained 99.8% uptime supporting 500K+ people accessing critical mental health services with zero major outages despite 4X traffic increase via auto-scaling infrastructure and rate limiting
- Optimized PostgreSQL queries with strategic indexes reducing query times from 8+ seconds to sub-200ms, added connection pooling decreasing page load time by 65%

WeWork | Miami, FL
Junior Backend Engineer
May 2018-January 2020 | On-Site

- Migrated Booking Service from Rails monolith to Java Spring Boot microservice with RESTful APIs and RabbitMQ achieving 85% latency reduction for sub-second booking confirmations across 400+ locations
- Deployed containerized services on AWS ECS with RDS Multi-AZ, CloudWatch monitoring, and Lambda-based reporting supporting 50K+ daily API requests at peak processing volumes
- Sustained 99.9% uptime for critical booking infrastructure during WeWork near-collapse with skeleton crew, preventing data loss and keeping revenue-generating systems operational across company restructuring
- Cut AWS infrastructure costs by 40% via rightsizing instances, S3 lifecycle policies, and RDS configuration optimization preserving performance SLAs across 18 months

Junior Frontend Engineer
July 2017-March 2018 | On-Site

- Rebuilt member dashboard with React/Redux featuring code splitting and lazy loading reducing initial page load time from 4.2s to 1.8s improving user experience and engagement metrics
- Created responsive mobile-first interface with optimistic UI patterns for space booking flows increasing booking completion rate by 34% via reduced friction

Education

Bachelor's Degree in Computer Science | Florida State University | *March 2013-May 2017*

Technical Skills

- **Programming Languages**: Python, Typescript/JavaScript, Java, Bash, SQL, React, Node.js
- **Frontend Development:** React, Vue.js, Angular, React Native, D3.js, Tailwind CSS, Bootstrap
- **Backend Development:** Node.js, Python, .NET Core, Java Spring Boot, Express.js, FastAPI, Flask
- **Cloud & DevOps:** AWS (EC2, ECS, Lambda, SageMaker), Azure (OpenAI, App Services), GCP (Cloud Run, BigQuery), Kubernetes, Docker, CI/CD, Terraform
- **Databases:** PostgreSQL, SQL Server, MongoDB, Redis, AWS Redshift, Pinecone, Weaviate
- **Data Engineering:** Apache Kafka, Apache Airflow, Apache Spark, ETL Pipelines, Stream Processing
- **Vector Databases & Search:** Pinecone, ChromaDB, FAISS, Weaviate, Milvus, pgvector
- **FinTech & Compliance:** Payment Processing (Stripe, Plaid), KYC/AML Systems, PCI-DSS Compliance, SOC 2, HIPAA, Fraud Detection, Risk Management
- **Development Tools & Practices:** Git/GitHub, VS Code, Postman, Prometheus, Grafana, Tableau, Jupyter, JIRA, API Design, Datadog, Code Review
- **LLM Architecture & Techniques:** RAG, Vector Search, Semantic Caching, Few-shot Learning, Chain of Thought Prompting, Agent Frameworks (LangGraph, AutoGPT)
- **AI/ML & LLM:** TensorFlow, PyTorch, scikit-learn, XGBoost, LangChain, LlamaIndex, OpenAI API, Anthropic Claude, Transformers, BERT, Prompt Engineering, Fine-tuning