Takumi Watanabe

Agentic AI & Cloud Consultant
w.takumi.cs@gmail.com | +1 (424) 999-5344 | Long Beach, CA | linkedin.com/in/takumi-watanabe
Available for C2C contracts | Remote or Los Angeles, CA | Rate: Negotiable based on project

PROFILE

AI Engineer with 13 years building scalable backend systems, currently specializing in production LLM and RAG systems using Python, LangChain, and AWS. Shipped Agentic RAG systems achieving 30% accuracy improvement and sub-second latency. Expert in agentic workflows, hybrid search, vector databases, Docker, and Kubernetes with full-stack ownership.

EDUCATION

Master of Science in Computer Science | California State University, Long Beach | December 2012
Bachelor of Science in Mathematics | California State University, Long Beach | May 2011

CERTIFICATIONS

AWS Certified Generative AI - Professional | 2026
AWS Certified Machine Learning - Specialty | 2026

TECHNICAL SKILLS

AI & Machine Learning:	LLM, RAG, Agentic Workflows, LangChain, Vector Databases (Qdrant), Ollama
LLM Evaluation:	RAGAS, LangSmith, Prompt Engineering, A/B Testing, LLM Tracing
Backend Development:	Python (FastAPI, Pydantic, Celery), SQL (PostgreSQL), REST APIs
Cloud & Infrastructure:	AWS (Lambda, ECS, S3, DynamoDB), Docker, Kubernetes, Terraform, CI/CD
Data Engineering:	Hybrid Search, Embeddings, BM25, RRF Fusion, Query Planning, HNSW

WORK EXPERIENCE

Principal Engineer
Stack Architect, Remote — July 2023 - Present

- Deployed enterprise RAG solution processing 10K+ documents with continuous evaluation and benchmarking, achieving 30% accuracy improvement over baseline and sub-second response times.
- Established gold-standard evaluation dataset using RAGAS metrics (Faithfulness, Answer Relevance) enabling data-driven deployments with quantifiable accuracy benchmarks.
- Architected agentic AI system with tool-calling, multi-turn reasoning, and context management enabling autonomous task completion and intelligent user interactions.
- Integrated local LLM inference with Ollama for privacy-first answer generation, eliminating external API costs while ensuring data security compliance for sensitive document workflows.

Principal Cloud Software Engineer
Cylance / BlackBerry, Irvine, CA — May 2019 - July 2023

- Built scalable Python microservices deployed to 6 AWS production regions serving 14M+ endpoints with 99.9% uptime for critical incidents and real-time security monitoring.
- Architected threat intelligence system processing 4M monthly event feeds using AWS Kinesis, Lambda, Kafka, and Elasticsearch for real-time security analytics serving enterprise customers globally.
- Reduced AWS infrastructure costs by 83% from $300K to $50K monthly ($250K annual savings) through event processing architecture redesign with cost-optimized data streams and auto-scaling.
- Automated infrastructure provisioning for 14 AWS resource types using Terraform IaC and Jenkins CI/CD reducing deployment time by 70% across 6 production regions.

Cloud Engineer Consultant

Microsoft (via Collabera), Remote, WA January 2022 - June 2022

- Modernized 2 legacy services refactoring 15 deeply nested object models into .NET Core microservices optimized for Kubernetes deployment with container orchestration and auto-scaling.
- Enhanced reliability and scalability of systems powering order fulfillment and financial transactions processing $50M+ annual revenue on Microsoft internal marketplace.

Principal Software Consultant

Bio-Rad (via Technossus), Irvine, CA January 2017 - May 2019

- Developed 6 microservice REST APIs using Python and Node.js in AWS reducing infrastructure costs by 40%.
- Implemented OAuth 2.0, JWT authentication, and RBAC across 15 microservices achieving SOC 2 compliance.

PROFESSIONAL PROJECTS

RAGnosis | https://www.ragnosis.app 2025 - Present

Production agentic RAG system with query planning, hybrid search, and automated quality evaluation

- Architected agentic RAG pipeline with query planning, semantic expansion, and automated evaluation achieving 0.91 faithfulness score via RAGAS framework with 28 test questions.
- Implemented hybrid search combining vector (pgvector HNSW) and keyword search (BM25) with RRF fusion (60/40 weighting) improving accuracy over naive semantic search.

Agentic Code Review 2025 - Present

Autonomous security analysis system using LangChain agentic patterns and tool-calling

- Built autonomous agentic security system using LangChain with LLM-powered reasoning detecting OWASP Top 10 vulnerabilities, hardcoded secrets, and insecure patterns.
- Reduced false positive rate by 60% through confidence scoring, context-aware validation, and iterative refinement loops with self-hosted Ollama and Docker deployment.

CORE COMPETENCIES

Full-Stack System Ownership • Data-Driven Decision Making • Cross-Functional Collaboration • Rapid Prototyping & Iteration • Performance & Cost Optimization • Technical Problem Solving • Production Reliability Engineering • Technical Mentorship & Knowledge Sharing • Continuous Learning & Innovation