

Richard Watson Stephen Amudha

Chantilly, VA | (413) 430-9201 | rstephenamud@umass.edu | github.com/RW2523 | linkedin.com/in/richardwatsonsa | richardwatsonsa.netlify.app

SUMMARY

AI and full-stack software engineer with 4+ years building production LLM, agentic AI, and web systems end to end. Ships across the stack — React/TypeScript frontends, Python and Java backends, and LLM-powered features grounded in RAG and vector retrieval — and has scaled services to 30M+ requests/day at sub-200 ms for 1M+ daily users. Owns problems from design and architecture through deployment and operations, and builds AI with the guardrails production needs: approval workflows, RBAC, and audit logging. MS in Computer Science (4.0 GPA).

EDUCATION

University of Massachusetts Amherst, Amherst, MA
Master of Science in Computer Science | GPA: 4.0/4.0

Jan 2025 – Dec 2026 (Expected)

Anna University, Chennai, India
B.E., Electronics and Communication Engineering | GPA: 3.5/4.0

Aug 2018 – Apr 2022

TECHNICAL SKILLS

Languages: Python, Java, TypeScript, JavaScript, SQL, C++

AI & LLMs: LLM Integration (AWS Bedrock, Claude, Llama, Hugging Face), Agentic AI, Function Calling & Tool Use, Prompt Engineering, RAG, GraphRAG, LangChain, LangGraph, PyTorch, Embeddings & Vector Search

Backend & APIs: FastAPI, Spring Boot, Node.js, RESTful API Design, Microservices, Distributed Systems, Event-Driven Architecture, JWT/OAuth, RBAC

Frontend: React.js, Next.js, HTML5, CSS3, Single-Page Applications (SPA), Responsive UI

Data & Cloud: PostgreSQL, MySQL, Redis, pgvector, Vector Databases (Pinecone, FAISS), AWS (EC2, S3, Lambda, Bedrock), Docker, Kubernetes, Apache Airflow

Practices: System Design, CI/CD (GitLab), Code Reviews, Automated Testing, Git, Agile/Scrum, Jira

PROFESSIONAL EXPERIENCE

Ajace | Software Developer Intern

Nov 2025 – Present

- Engineered **EchoMind**, a production AI platform for transcription, semantic search, and RAG-based LLM querying, as sole engineer on the codebase — owning design, implementation, and deployment across cloud, mobile, and edge.
- Developed full-stack LLM workflows (document ingestion, embeddings, vector retrieval, grounded response generation) exposed through RESTful APIs, and prototyped an AWS Bedrock RAG capability on foundation models.
- Architected a Kubernetes-orchestrated, event-driven data pipeline (Apache Airflow, RabbitMQ, Redis) for large-scale ingestion, parsing, extraction, and validation.
- Ported Transformer inference from x86 to ARM, improving throughput ~25% through operator fusion and batching.

CodeWorks | Software Engineering Intern (AI/ML)

May 2025 – Aug 2025

- Built an OCR and extraction pipeline (Tesseract, EasyOCR) converting 1,000+ unstructured inputs into structured data at 85% accuracy.
- Fine-tuned detection models to 92% mAP and developed the evaluation harness that gated every release — precision 0.88, recall 0.84, F1 0.86.

Zoho Corporation | Member Technical Staff

Jan 2022 – Jan 2025

- Led development of **Ask ZIA**, an in-app AI assistant used daily by 1M+ people — owning it from architecture and design decisions through production, building the LLM-backed REST APIs and the React frontend.
- Engineered distributed backend services scaled to 30M+ requests/day at sub-200 ms latency, improving model accuracy 15% via PyTorch/Hugging Face Transformer systems.
- Refactored legacy JavaScript/JSP into clean React UI, eliminating 20K+ lines of redundant code and improving frontend performance by 75%.
- Delivered containerized AI microservices adopted across 30+ products, and automated CI/CD pipelines that cut deployment time 40%.

PROJECTS

AM Meeting | Agentic AI Assistant (React, FastAPI, RAG, Function Calling) | github.com/RW2523/ammeet

- Built an AI agent that grounds reasoning in retrieved context via pgvector RAG, then takes actions through function calls into Slack, email, and Jira — every action gated by approval controls, RBAC, and audit logging.

IntelliGraphRAG | GraphRAG Platform (LangChain Orchestration) | github.com/RW2523/intelligraphrag

- Architected a three-layer GraphRAG platform unifying structured and unstructured sources into a knowledge graph, with LLM reasoning over combined vector and graph retrieval.

Zenalam | Full-Stack Web Application (React, Spring Boot, PostgreSQL) | github.com/RW2523/wellness1

- Created a React and Spring Boot application analyzing 5,000+ records into personalized insights, with Random Forest models and PostgreSQL-backed storage powering user-facing dashboards.