

Mehdi Salhi

AI & Cloud Engineer

mehdisalhi.dev@gmail.com | (718) 500-1191 | Stamford, CT | U.S. Permanent Resident (no sponsorship needed)

[linkedin.com/in/mehdi-salhi-work](https://www.linkedin.com/in/mehdi-salhi-work) | github.com/medysaly | mehdisalhi.com

SUMMARY

AI and Cloud Engineer and AWS Certified Solutions Architect Associate building AI-powered systems on AWS. Computer Science student at SNHU (graduating August 2026, GPA 3.71) with hands-on experience integrating large language models into production applications, building retrieval-augmented generation (RAG) pipelines, and designing agentic tool-calling workflows on Amazon Bedrock, deployed on serverless AWS infrastructure. Skilled in Python, LLM integration, RAG, prompt engineering, and infrastructure as code. U.S. permanent resident, no sponsorship needed.

TECHNICAL SKILLS

AI & ML: Amazon Bedrock, Large Language Models (LLMs), RAG, Hybrid Search, Cross-Encoder Reranking, Agentic / Tool-Calling Workflows, Prompt Engineering, LLM Evaluation, Amazon SageMaker

Cloud (AWS): Lambda, API Gateway, DynamoDB, S3, EC2, IAM, EventBridge, CloudWatch, Bedrock

Infrastructure & Deployment: Terraform, AWS SAM, Docker, GitHub Actions, CI/CD, OIDC

Languages: Python, Bash, TypeScript/JavaScript, SQL, YAML

PROJECTS

Unkommon — Serverless AI platform on AWS | Amazon Bedrock, LLMs, Python, Lambda, AWS SAM | github.com/medysaly/unkommon · unkommon.ai

- Designed and deployed a serverless AI platform on AWS running a website chatbot and a voice agent on large language models via Amazon Bedrock.
- Built agentic, tool-calling workflows where the assistant checks availability and books appointments through function calls, using Python and Bedrock.
- Engineered the system prompt, tool schemas, and guardrails (including prompt-injection defenses), and added a Trie-based intent classifier to answer common questions before the LLM, cutting latency and per-call cost.
- Deployed on serverless AWS (Lambda, API Gateway, DynamoDB) as infrastructure as code with AWS SAM, with a CI pipeline and automated tests.

Company Policy RAG — Retrieval-Augmented Generation system | Python, RAG, Docker, FastAPI | github.com/medysaly/company-policy-rag

- Built a RAG pipeline with hybrid retrieval (dense embeddings + BM25) and cross-encoder reranking over proprietary documents.
- Evaluated retrieval quality with faithfulness and context precision, both 1.00 on a 10-question test set.
- Containerized with Docker and deployed with a FastAPI backend on Hugging Face Spaces.

AWS Cost Watchdog — Serverless FinOps tool on AWS | Terraform, Lambda, EventBridge, Python | github.com/medysaly/aws-cost-watchdog

- Built a serverless tool that uses ML-based anomaly detection to flag unusual AWS spend, with idle-resource detection and tag governance, surfaced in a React dashboard.
- Provisioned with Terraform (remote S3 state) and deployed via a GitHub Actions CI/CD pipeline using OIDC; four event-driven Lambda functions on EventBridge.

CERTIFICATIONS

AWS Certified Solutions Architect – Associate (2026) • AWS Certified Cloud Practitioner (2026) • Google IT Support Professional Certificate (2026)

EDUCATION

B.S. in Computer Science, Southern New Hampshire University — Expected August 2026 • GPA 3.71