Daniel Romero

LinkedIn: linkedin.com/in/infoslack | infoslack.pro | infoslack@gmail.com | GitHub | +5585997116130

INTRO

AI Engineer with 25 years in tech and hands-on experience inside OpenAI, contributing to a project acquired by the company. I operate across the full stack of intelligent systems, from GPU-level kernel optimization with Triton to production RAG platforms handling 70M+ documents and end-to-end agentic workflows. I don't just prototype, I take LLM systems from research to production, with deep expertise in inference optimization, hybrid search, and scalable backend architecture.

EXPERIENCE

Stealth AI Startup (Acquired by OpenAI) AI Engineer – *Remote, USA 07/2025 – 02/2026*

- Contributed to large-scale LLM workloads at a stealth AI startup later acquired by OpenAI, with pipelines integrated directly into OpenAI production systems.
- Following the acquisition, continued working on the acquired project within OpenAI, developed high-performance Triton kernels integrated into PyTorch extensions, reducing GPU inference latency through kernel-level optimization.
- Profiled GPU bottlenecks with Nsight and PyTorch Profiler; applied mixed-precision training (FP16/BF16) and INT8 quantization to improve throughput and reduce memory footprint.
- Designed agentic workflows with LangChain and LlamaIndex; implemented hybrid search pipelines with Qdrant; built backend APIs with FastAPI to serve RAG pipelines and agent execution.

Tech Stack: Python, PyTorch, Triton, CUDA, LangChain, LlamaIndex, Qdrant, FastAPI

Doximity Senior Software Engineer / AI – *Remote, USA 07/2024 – 07/2025*

- Designed and built a production-grade RAG platform enabling semantic search over 70M+ medical documents, cutting retrieval time for healthcare professionals by delivering more accurate results at scale.
- Implemented a hybrid search pipeline combining dense, sparse, and late interaction embeddings with re-ranking, significantly improving retrieval precision over baseline keyword search.
- Built a FastAPI backend with dedicated endpoints for embedding generation, document ingestion, and vector search, integrated with a Qdrant vector database.

Tech Stack: Python, FastAPI, Qdrant, HuggingFace, OpenAI, LangChain, Docker, Kubernetes

Qdrant ML Engineer / AI – *Remote, Germany 01/2024 – 07/2024*

- Produced technical tutorials, documentation, and videos covering semantic search, vector search, and RAG, helping developers onboard faster onto the Qdrant ecosystem.
- Contributed to framework integrations including DSPy, LangChain, and LlamaIndex, enabling smoother adoption of Qdrant in production retrieval pipelines.
- Served as internal consultant (under NDA) for enterprise customers integrating Qdrant into large-scale production retrieval systems, accelerating adoption and reducing integration friction.

Tech Stack: Python, LangChain, LlamaIndex, DSPy, Qdrant, OpenAI, HuggingFace

DNSFilter Platform Engineer / ML – *Remote, USA 06/2022 – 12/2023*

- Reduced application deployment time by 87%, from 5 minutes to 40 seconds by migrating Ruby on Rails services from bare-metal infrastructure to Kubernetes (EKS) using Terraform and Ansible.
- Improved incident response time across production systems by implementing continuous alerting and observability using Loki, Grafana, Datadog, and eBPF, eliminating blind spots in the monitoring stack.

- Collaborated with the data science team to build ML training pipelines in PyTorch, supporting model development on scalable cloud infrastructure.

  Tech Stack: Terraform, Ansible, Docker, Kubernetes (EKS), Loki, Grafana, Datadog, eBPF, PyTorch

Henku Data Scientist / ML – *Remote, Estonia 04/2021 – 06/2022*

- Reduced model training and deployment time by 90%, from 30 minutes to 3 minutes, by replacing ad-hoc Jupyter notebooks with a continuous training pipeline using transfer learning, serving 10K+ active users in Spain.
- Built a computer vision model from scratch using a CNN in PyTorch to identify the nutritional value of food, taking it from prototype to production with 10K+ active users.
- Automated the model development lifecycle using Airflow, Kubeflow, and Weights & Biases, eliminating manual steps and improving reproducibility across experiments.

  Tech Stack: Python, PyTorch, TensorFlow, Scikit-Learn, Airflow, Kubeflow, Weights & Biases

Elastic Senior Site Reliability Engineer – *Remote, USA 09/2018 – 04/2021*

- Maintained 99.95–99.99% uptime across tens of thousands of Elasticsearch clusters on AWS, Azure, and GCP by leading on-call incident management in a multi-region, multi-cloud environment.
- Contributed to a Golang CLI tool for Kubernetes cluster maintenance that eliminated ~8 hours of manual operational work per week for the SRE team.
- Expanded Elastic Cloud's global footprint by deploying new regional infrastructure, increasing platform availability and reducing latency for international customers.

  Tech Stack: Golang, Python, Ansible, Terraform, Kubernetes, AWS, Azure, GCP, Elasticsearch

OTHER EXPERIENCES

- DevOps Engineer (2015–2018): AWS, GCP, Docker, Kubernetes, Ansible, Terraform, Prometheus, Grafana
- Software Developer (2007–2015): PHP, Java, Ruby on Rails, MySQL, PostgreSQL, LXC, Docker
- Sysadmin (2000–2007): Linux, Iptables, proxies, networks, Bash, Perl