

Dario Fumarola

New York, NY, USA • dariofumarola90@gmail.com • +1 540-466-2013 • fumadari.github.io

RESEARCH INTERESTS

Reinforcement learning for training and steering LLM and embodied agents; adaptive computation and efficient inference; runtime safety and behavioral control; multi-agent interaction and social inference; evaluation of agent behavior and training dynamics; large-scale vector search and retrieval.

EDUCATION

Washington and Lee University

Bachelor's Degree in Computer Science and Mathematics

Lexington, VA

2019 – 2023

- **Honors:** Davis United World College Scholar.
- **Relevant coursework:** Deep Learning, Real Analysis, Differential Geometry, Probability, Social Networks.

PROFESSIONAL EXPERIENCE

Solutions Engineer, AI Prototyping

New York, NY

Amazon Web Services

2023 – Present

- Technical lead for 20+ enterprise AI engagements, from discovery with customer leadership through production deployment, owning architecture, implementation, and evaluation; led teams of 5+ engineers on concurrent 1–4 week production builds for pharmaceutical, automotive, and other enterprise customers.
- Built a multi-agent system that deploys protein models for pharmaceutical lab scientists, automating packaging, GPU scheduling, provisioning, validation, observability, and recovery with human approval at key steps; iterated on prompts, tools, and evals, cutting deployment time from weeks to hours; adopted by 10+ customers.
- Codified an evaluation-driven deployment framework spanning representative test sets, deterministic checks, model graders, human review, and production telemetry; reusable templates adopted across AWS Prototyping cut repeat prototype-to-production time by 30%.
- Diagnosed and fixed inference bottlenecks in production Amazon Bedrock workloads using dynamic batching, request scheduling, streaming, and custom CUDA and Trainium NKI kernels; reduced time to first token by 35% and increased throughput $1.8\times$, with the optimizations upstreamed into Bedrock.
- Partnered with an autonomous-driving customer's perception team to solve a training-data bottleneck; built a diffusion-based camera-to-radar synthesis pipeline from archived footage, avoiding 50K+ miles of new data collection and raising object detection accuracy by 12% relative to the prior model.

PRESENTATIONS & MANUSCRIPTS

Fumarola, D., Tan Ruan, J. *Knobs of the Mind: Dopamine, Serotonin, and a Maze-Running Rover.*

International Conference on Machine Learning, **ICML Expo**, 2025.

Fumarola, D., et al. *Who Needs Attention Anyway? Elastic State Models for Real-Time Streaming Tasks.*

Neural Information Processing Systems, **NeurIPS Expo**, 2025.

Fumarola, D., Gaikwad, A. *Broadcast-Gain: A 2-Byte, Stop-Gradient Control Plane to Trim Long-Tail Latency.*

New York Reinforcement Learning Workshop, **NY-RL**, 2025.

La Malfa, E., **Fumarola, D.** *Pretraining for Deliberation: Reinforcement Learning of Internal Computation in LLMs.*

Manuscript in preparation.

PATENTS

Fumarola, D., Ferhatosmanoglu, H., Wagner, T., Warfield, A. *Selective Vector Replication for Nearest Neighbor Search.*

P89832-US01, filed 2025. The method is used in Amazon S3 Vectors, which serves tens of billions of vectors.

RESEARCH EXPERIENCE

Machine Learning Researcher

New York, NY

Amazon Science

2023 – Present

- Research rotation with Amazon Scholars on billion-scale similarity search for Amazon S3 Vectors, analyzing HNSW neighborhood structure and developing replication strategies.
- Technical co-lead on a \$500K Amazon Research Award project under Build on Trainium, led by Emanuele La Malfa, University of Oxford, on safe, social pretraining of LLM agents.
- Research on neuromodulated RL and adaptive-compute mechanisms for agent control, leading large-scale experiments, analysis, and writing.

Undergraduate Researcher, Mathematics

Lexington, VA

Washington and Lee University

2022 – 2023

- Studied numerical methods for nonlinear ordinary differential equations in collaboration with mathematics faculty, with emphasis on stability, stiffness, and eigenvalue-based analysis.
- Implemented MATLAB experiments comparing time-stepping schemes and visualizing phase portraits and stability regions for representative dynamical systems.
- Produced write-ups and reproducible scripts that let other students and faculty replicate the experiments and reuse the visualizations in advanced coursework.

SELECTED RESEARCH PROJECTS

Safe, Social Pretraining of LLM Agents

2026 – Present

Investigating whether reinforcement-learning pretraining on social interactions improves agents' cooperative reasoning and safety; built the training and evaluation stack that tracks both across checkpoints on Trainium.

Elastic State Models: Bounded Adaptive Inference

2025

Built an inference-time adaptation layer for a frozen streaming state-space model under a fixed compute budget; raised grid-maze navigation success from 68% to 85%, halved collisions, and cut constraint violations by 64%.

Mood Manifolds: Runtime Safety & Behavioral Steering

2025

Developed a runtime control layer of neuromodulatory “mood knobs” that steers embodied agents from live hazard signals without retraining; one policy covered most of the reward–safety frontier of 64 separately trained specialists.

Broadcast-Gain Overlay for Cooperative Multi-Agent RL

2025

Developed a 2-byte, stop-gradient broadcast overlay for PPO+GAE agents in congested junction grids with packet loss, reducing p99 traversal latency by 15% versus a no-broadcast baseline under matched traffic conditions.

Selective Replication for Sharded Vector Search

2024

Co-invented a selective-replication method for sharded vector search and built a hierarchically partitioned prototype on Amazon S3; reduced recall error by 28% at a fixed vector-access budget. It is used in Amazon S3 Vectors.

TEACHING & MENTORING

Course Assistant, MATH 332: Ordinary Differential Equations

Lexington, VA

Washington and Lee University

2021

- Authored comprehensive $\text{\LaTeX}$ solution guides and review materials covering existence and uniqueness, boundary value problems, and numerical integration schemes.
- Held weekly office hours and exam reviews and supported MATLAB modeling assignments, guiding 30+ students in implementing numerical solvers and interpreting results.

TECHNICAL SKILLS

AI / LLM: Agents and tool use, evaluation harnesses, RAG, LLM inference/serving, RL, multimodal, vector search.

ML & Infra: PyTorch, JAX, CUDA, Trainium/NKI, Neuron/XLA, distributed training, Kubernetes, Redis.

Software: Python, C++, TypeScript/JavaScript, MATLAB, FastAPI, React, CDK, Terraform.