

Aryan Srivastava

aryan7srivastava@gmail.com | linkedin.com/in/aryan-srivastava-20503b2ab | aryan-srivastava.com

SUMMARY

Applied AI engineer who takes AI systems from idea to production for enterprise clients: agentic systems, LLM fine-tuning and serving, RAG, and forecasting. Owns the full solution, from multi-GPU training to the API, the interface, and the security layer, for clients in automotive, retail, and enterprise networking. Mathematics and economics background; focused on measured results.

PROFESSIONAL EXPERIENCE

Velocis Systems (joined through its subsidiary Agilus Technologies) — *AI Engineer*

Sep 2025 – Present

Velocis Foresight — AI Spares-Forecasting Platform

- Built demand forecasting for a leading automotive OEM over 21,957 spare-part SKUs and 52 months of history: a LoRA adapter (3.6% of parameters, ~7 min on one H200) on Cisco's time-series foundation model.
- Cut pooled forecast error (WAPE) from 17.2% to 14.9% and raised forecast value-add over the naive method from 18.5% to 29.5%, after a benchmark against 15 statistical and deep-learning models.
- Shipped the product end to end: a planner console, an order-quantity engine on demand quantiles, a Webex and in-console assistant with a grounding check that blocks figures no tool returned, Cisco Duo and AI Defense security, and Splunk drift alerts.

Bullseye Zero — Agentic Root-Cause Analysis for Enterprise Networks

- Designed and built a multi-turn RCA agent that turns a user complaint into a cited report: the LLM picks the next check, and deterministic code computes device state, cause ranking, and confidence over a network dependency graph.
- Wrote read-only connectors for Cisco Catalyst Center and Cisco ISE behind ~24 vendor-neutral tools, with an append-only MongoDB evidence store, FastAPI, a React UI, and 98 tests.

Cybersecurity LLM Fine-Tuning (Continual Pre-Training, SFT, GRPO)

- Built the training pipeline that aligns Qwen2.5-7B to cyber threat-intelligence tasks on 6×H200 GPUs with DeepSpeed ZeRO-2, with a 72B LLM-as-judge reward model served on vLLM.
- Designed task-specific reward functions and difficulty-based data filters across four CTIBench benchmarks, which cut zero-variance training steps from 80% to under 30%.

LLM Serving and Generative AI Solutions

- Deployed production LLMs (Qwen-72B, Foundation-Sec-8B) on Red Hat OpenShift with vLLM tensor parallelism, GPU-aware scheduling, and autoscaling; red-teamed the endpoints with Cisco AI Defense.
- Delivered client solutions on NVIDIA NIM: a multimodal RAG shopping assistant with NemoGuard guardrails, catalog enrichment (text, images, 3D models), and document and video Q&A.
- Built a LangGraph network-compliance agent for ~10,000 managed devices (MongoDB, Qdrant, Redis cache) and a real-time voice agent over WebSocket audio streams.

CoreOps.ai — Data Scientist

Sep 2024 – Sep 2025

- Core developer of AgentCORE, an enterprise platform to build and deploy multi-agent systems across SAP S/4HANA, ERP, and CRM.
- Built the ML lifecycle tools of the platform — data ingestion, training, experiment tracking, model registry, deployment, and live monitoring — with a web UI and an open-source CLI, plus CI/CD and role-based access control.

Mphasis — AI/ML Intern, client: Hewlett Packard Enterprise

Jun 2022 – Sep 2022

- Built healthcare AI proofs of concept: medical text classification (fine-tuned DistilBERT, F1 0.87), sentiment analysis, and a nurse-staffing recommender, delivered as Flask REST APIs.

FOUNDER AND INDEPENDENT PROJECTS

PrepVue (prepvue.ai) — Founder

2025 – Present

- Built an AI voice app for interview practice: a real-time Deepgram → GPT-5 → ElevenLabs pipeline over WebSockets and 15 drill-specific LLM evaluators with strict JSON-schema outputs, on FastAPI and PostgreSQL.

montecarlo-zero, alphasft — Game AI research

2022 – Present

- An AlphaZero-style chess engine in PyTorch (ResNet policy and value heads, MCTS self-play), and a Transformer board-strength model for Teamfight Tactics trained on placement data from the Riot API.

SKILLS

AI and ML: PyTorch, Hugging Face, TRL, scikit-learn, LLM fine-tuning (SFT, LoRA, GRPO, DPO), RAG, agentic AI (LangGraph, tool use), time-series forecasting, evaluation

Infrastructure: vLLM, NVIDIA NIM, DeepSpeed, Docker, Kubernetes, OpenShift, multi-GPU training, CI/CD, MLflow, Weights & Biases

Engineering, data, security: Python, SQL, FastAPI, React, MongoDB, PostgreSQL, Redis, Qdrant, Milvus, Splunk, Cisco AI Defense, Cisco Duo, NeMo Guardrails

EDUCATION

Boston University — Bachelor of Arts in Economics and Mathematics (Probability, Statistics, Time Series, Optimization) *May 2026*