Sawyer Czupka

+1 (757) 818-0742 — sawyerczupka@gmail.com — sawyerczupka.com — linkedin.com/in/sawyerczupka

ML engineer with 2+ years designing and deploying production RAG systems, LLM pipelines, and AI evaluation infrastructure across startup, research, and enterprise environments — with hands-on model training across CNNs, Transformers, GNNs, and diffusion architectures, and deep familiarity with LLM fine-tuning workflows from retrieval system design to cloud deployment.

SKILLS

ML & AI Technologies
Python, TypeScript, Rust, PyTorch, Keras, Scikit-Learn, Hugging Face Transformers, LoRA/QLoRA, Axolotl, LlamaIndex, Qdrant, DeepEval, vLLM, Azure OpenAI, AWS Bedrock, RAGAS, FastAPI, PostgreSQL, Git

Infrastructure & Engineering
Docker, Kubernetes, Terraform, ArgoCD, GitHub Actions, CI/CD, Azure, AWS Lambda, API Gateway, Modal, Prometheus, Grafana, MLflow, Linux, React, Vue

EDUCATION

College of William & Mary *Sept. 2021 – May 2025*
Bachelor of Science (B.S.) in Computer Science & Data Science, Cum Laude & Dean's List GPA: 3.6
Relevant Coursework: Neural Networks, Natural Language Processing, Cloud Computing, Cybersecurity

PROFESSIONAL EXPERIENCE

Software Engineering Intern, Luminexis AI / Threat Tec *Aug. 2025 – Dec. 2025*

- Served as sole engineer owning the entire stack; built production React/FastAPI platform with JWT auth and RBAC while owning all AI system design, evaluation, and deployment decisions end-to-end.
- Deployed production RAG chatbot with Qdrant vector database and LlamaIndex achieving 70–80% on RAGAS evaluation metrics; built automated DeepEval test suite with corpus-derived test cases for continuous quality measurement across configurable knowledge bases.
- Designed multi-tenant LoRA fine-tuning architecture for per-customer model specialization: planned hierarchy of customer specific LoRA adapters on top of a shared base model, served via runtime adapter-swapping in vLLM; selected Axolotl for training orchestration and Modal for on-demand GPU provisioning to avoid prohibitive Azure GPU costs, with full LoRA vs. full-rank tradeoff analysis documented for team review.
- Engineered LLM-powered document analysis pipeline using Azure OpenAI and Document Intelligence to automatically extract requirements, risk factors, and action items from 140+ contract PDFs, reducing manual review from hours to minutes.

Machine Learning & Software Engineer, Teamculture.ai / L10.tech *Jan. 2024 – Sept. 2024*

- Spearheaded serverless RAG evaluation system on AWS Lambda and API Gateway using AWS Bedrock as the LLM endpoint and DeepEval as the evaluation framework; modular architecture enabled automated, scalable AI performance monitoring.
- Built Vue/FastAPI interface for evaluation management and golden dataset curation; hybrid strategy combining auto-generated and manually curated test cases enabled systematic, continuous AI quality improvement over time.

Machine Learning Technical Lead, GeoLab @ William & Mary *Jan. 2023 – May 2025*

- Led ML integration across frontend, backend, and ML subteams for SCOPE research platform; architected microservice RAG system on Kubernetes using Qdrant, vLLM, LlamaIndex, and FastAPI with independent scaling of model inference from application logic.
- Built VLM-powered document processing pipeline using open source models (Qwen-2.5-VL and OlmOCR) achieving quality comparable to frontier models (empirically validated against Claude as baseline) at a fraction of the cost; 50%+ reduction in parsing errors vs. conventional non-ai tools.

PERSONAL PROJECTS

LapEvo: iRacing Telemetry Tracker & Racing Coach *Jan. 2025 – Present*

- Architected real-time 60Hz telemetry pipeline (Rust client, FastAPI, TimescaleDB) with heuristic lap analysis engine comparing braking zones, brake points, and corner shapes against a reference lap to deliver pre-corner coaching cues; designing dataset collection strategy to train a future ML coaching model.

Personal Homelab Infrastructure *July 2023 – Present*

- Operate 116-pod Kubernetes cluster on Proxmox with full IaC stack (Packer, Terraform, ArgoCD GitOps) and Prometheus/Grafana observability; production-like environment for experimentation with ML serving infrastructure (vLLM, model APIs).

SELECTED ML PROJECTS

William & Mary *2021–2025*

- Molecular property prediction with GNNs (PyTorch Geometric); transformer-based sentiment analysis (Keras); CNN image classification & regression; ML-based network traffic classifier (XGBoost/Scikit-Learn); time series forecasting; generative modeling with diffusion models (guided, CFG).