

Kyler Russell

Lenexa, KS · 913-337-6518 · KylerRssll@gmail.com · LinkedIn: Kyler-Russell · GitHub: KylerRussell

SUMMARY

Computer Science student at the University of Kansas and independent machine learning researcher. Runs long-term solo research projects in extreme low-bit LLM quantization, recurrent and brain-inspired architectures, and reinforcement learning on a self-built 2× RTX 3090 / ~629 GB RAM server. Works hypothesis-first — paired controls, multiple seeds, falsification tests — and documents negative results as carefully as positive ones. Also brings network security (CyberPatriot, 5th nationally of 5,266 teams), Kubernetes infrastructure, and entrepreneurship (100,000+ units sold).

TECHNICAL SKILLS

Languages: Python, Java, C++, Bash/Shell, SQL, JavaScript

Machine learning: PyTorch (custom autograd functions, DDP, mixed precision, torch.compile/Triton), Hugging Face datasets/tokenizers, Ray, lm-evaluation-harness, knowledge distillation, quantization-aware training, RL (policy gradients, GAE)

LLM compression: GPTQ, QuaRot/Hadamard rotation, block-wise reconstruction, straight-through-estimator QAT, ternary and sub-2-bit weight formats, llama.cpp / GGUF / ggml

Performance engineering: GPU memory profiling, pinned-memory and asynchronous host-to-device transfer, CPU optimizer offload, multi-GPU training on consumer hardware

Infrastructure: Kubernetes (Talos Linux), Argo CD / GitOps, Kustomize, Docker, GitHub Actions, Prometheus/Grafana, Longhorn, MetalLB, Tailscale, Sealed Secrets, NVIDIA device plugin, Linux administration

Security: Network security, firewall configuration, vulnerability remediation, system hardening

EDUCATION

University of Kansas — B.S. Computer Science

Aug 2024 – May 2027 (expected)

Korea University, Seoul — Study Abroad, Computer Science

Aug – Dec 2025

Relevant coursework: Artificial Intelligence, Deep Learning, Databases

RESEARCH & PROJECTS

Ternary LLM Quantization for llama.cpp

May 2026 – Present

Model-to-Ternary · Python, PyTorch, llama.cpp/GGUF

- Building an end-to-end pipeline to compress Qwen3.6-27B (64-layer hybrid Gated-DeltaNet / full-attention model) to ternary $\{-1, 0, +1\}$ weights for llama.cpp deployment: Hadamard rotation → calibration → block-wise reconstruction → latent assignment training → end-to-end distillation → GGUF export; iterating on a Qwen 4B testbed.
- Designed TQ1_64, a self-contained 1.78-bit-per-weight ternary format with a bit-exact Python reference encoder; finer 64-weight scale groups raised held-out token agreement from 79.3% to 82.1% over stock 256-weight formats.
- Replaced SGD-based layer reconstruction with a closed-form Gram-matrix solve, cutting compute ~20–30×.
- Diagnosed 3.5–4× training slowdowns as serialization in the CPU-optimizer path, not bandwidth; fixed with a multi-threaded fused Adam, pinned ring buffers for async host-to-device copies, and a custom autograd function.
- Fit a measured memory model (18.9 bytes per trainable latent, accurate to 0.2 GB) to plan 27B-scale full-latent training; an 8-bit CPU Adam roughly halves memory while matching 99.85% of fp32 update decisions.
- Wrote a measured survey of sub-2-bit weight formats across 13 structural families on stock Qwen 4B/27B weights; found reconstruction error predicts perplexity almost perfectly ($r = 0.996$) and reported where the project's own format was Pareto-dominated.
- Diagnosed free-generation failures (looping, non-termination inside reasoning traces) as exposure bias; built gated evaluations for commit rate, loop rate, compression ratio, and GSM8K correctness.

Continuous Thought Machine Language Models

Apr 2026 – Present

CTM-Transformer · Python, PyTorch

- Running a controlled study of whether temporal synchronization (Continuous Thought Machines, arXiv:2505.05522) improves recurrent computation in language models, against standard Transformer and recurrent-depth Transformer (arXiv:2502.05171) baselines trained on identical data.
- Wrote a pre-registered research plan covering FLOP- and token-matched comparisons, 2×2 mechanism ablations, paired-bootstrap statistics, and staged go/no-go gates.
- Audited the codebase for correctness, finding silent issues such as permutation-invariant token attention, unwired configuration flags, and evaluation-harness tokenization errors.
- Found that at 400M training tokens, a shallow 16-tick CTM beats its matched 1-tick control by 0.157 nats; cross-position ticks added 18.6 points on a synthetic word-problem task.

- Ran tick-count controls showing extra iterations did not help at 100M tokens, and rejected several plausible mechanism hypotheses with targeted probes.

Living Brain: Biologically Grounded Neural Engine

Jan 2026 – Present

Living-Brain · Python, PyTorch, Ray

- Built a GPU predictive-coding engine of thousands of model neurons on hierarchical modular graphs, with NMDA/AMPA dynamics, shunting inhibition, interneuron subtypes, and short- and long-term plasticity.
- Developed a cerebellar module that passes a 26-benchmark neurobiological diagnostic suite, including eyeblink conditioning, VOR and saccade adaptation, forward/inverse models, and interval timing.
- Trained language learning through a developmental curriculum (holophrases → slot-and-frame → complex constructions), reaching mid-90s accuracy on the first stage; distributed experiments with Ray.

AlphaSlayer: RL Agent for Slay the Spire 2

2026

Python, PyTorch

- Designed a two-network agent: a combat network (~5–15M parameters) and a run-level network (~20–50M) with Transformer encoders, set-based card tokenization, and an action-embedding scoring head.
- Showed the bottleneck was run-level policy (HP and deck management before bosses), not combat skill, because scripted heuristics beat the trained network on boss-reach rate; redesigned training around GAE and paired-seed evaluation.

Home Kubernetes Cluster (GitOps)

Oct 2025 – Present

k8s-gitops · Talos Linux, Argo CD, Kustomize

- Operate a bare-metal Talos Linux cluster managed entirely from Git with Argo CD's app-of-apps pattern and self-healing sync, running 20+ services across 278 commits.
- Services include a Harbor registry, Prometheus/Grafana monitoring, Longhorn storage, MetalLB load balancing, Tailscale VPN access, Sealed Secrets, and GPU workloads (Ollama, GPU-backed remote dev desktops) via the NVIDIA device plugin.
- Built a GitHub Actions pipeline that builds and pushes Docker images for Ray-based ML training.

Linux Desktop Environment

Jun 2026

personal-dotfiles · Bash, Python

- Modular Hyprland, Kitty, Waybar, and Rofi configuration for virtualized environments, with one-command install/update scripts and a custom Python telemetry HUD.

Website Redesign — Heart of America Relocation Council

Oct 2024

harckc.org · Wild Apricot, HTML/CSS, JavaScript

- Carried out a graphic design overhaul of the outdated website for a Kansas City 501(c)(6) nonprofit, building a custom theme on the Wild Apricot membership platform: page templates, layouts, typography, color system, and widget styles.

EXPERIENCE

Independent ML Researcher & Home Lab Administrator — Self-directed

2023 – Present

- Design, build, and maintain the compute and software stack behind the research above: a multi-GPU Linux server, a Kubernetes cluster, and networked services secured with firewall rules, access controls, and VPN-only exposure.
- Manage projects with Git, CI/CD, research plans, and results logs; each experiment records its hypothesis, controls, and outcome.

Programming Team Leader — FIRST Tech Challenge (FTC)

2022 – 2023

- Led development of Java robot control software, including autonomous image detection, driver-assist systems, and real-time sensor integration.
- Ran source control and code reviews, and debugged hardware-software integration under competition conditions.
- Took features from prototype through testing to competition deployment.

Networking Specialist & Club President — CyberPatriot

2021 – 2024

- Placed 5th nationally out of 5,266 teams in the national youth cyber defense competition.
- Secured operating systems and network configurations under strict time limits: firewall policy, vulnerability remediation, and system hardening.
- Served as club president while competing as the team's networking specialist.

Co-Founder & Co-CEO — ComfyCup LLC

2016 – Present

- Co-invented and scaled a consumer product to 100,000+ units sold worldwide.
- Coordinate design, manufacturing, and marketing as a small founding team.

LEADERSHIP & HONORS

Eagle Scout (BSA Troop 587) · **CyberPatriot National Finalist** (5th of 5,266 teams) · **German Silver Seal of Biliteracy** (2024)