

Dylan Jayabahu

☎ +1 (226) 505-0703 | ✉ dylan.jayabahu@uwaterloo.ca | 🌐 Github | 🔗 LinkedIn | 🌐 Website

EXPERIENCE

Research Intern, Mechanistic Interpretability | Google

Incoming · Jan – May 2027

ML Engineering Intern | Shopify

May – Aug 2026

- Rebuilt a support-triage model as a **chat-segmented ModernBERT classifier**; reduced healthy-chat false alarms from **89% to 17%** in offline trajectory evaluation at approximately 90% escalation recall.
- Trained on **~2.4M segments across 8 H200 GPUs** (DDP); reconciled offline and production evaluation (logged-score $r = 1.0$), then shipped a separate **shadow endpoint** with single-flag cutover/rollback.
- Recalibrated dashboard thresholds on approximately **1,900 shadow-scored conversations**; live false alarms were **41% vs. 18%** for the deployed and new models at comparable recall.
- Built an interpretability layer (occlusion + streaming- Δ attribution) and ran a **leakage ablation**: AP remained 0.960 after escalation language was removed (0.994 before removal).

Embedded Systems & Software Engineering Intern | Perpion

June – Sept 2025

- Reduced **C++ service latency by 12%** through manual memory management and cache alignment.
- Co-authored **3+ provisional patents** for hardware-software co-design addressing interrupt latency and determinism.

AI Development & Research Intern | Chirp

July – Sept 2024

- Built **C++/CUDA inference engines** for human action detection on edge devices, achieving **8× throughput** and **+5% F-score**.
- Tuned model architecture and quantization for **millisecond-latency inference**.

EDUCATION

University of Waterloo, Bachelor of Computer Science

GPA: **95/100**

RESEARCH & PUBLICATIONS

The Halt Vector: Internalizing a Causal Steering Intervention for Efficient Reasoning

First author

COLM 2026 Workshop on Efficient Reasoning · **Spotlight** 🔗

- Identified a causal stopping direction in a 7B reasoning model; steering controlled generation length ($\rho = -0.82$).
- Installed hook-free self-halting via full-activation reconstruction: **24% less thinking at held accuracy** on five unseen benchmarks, from 24 problems without RL; reductions tracked removable reasoning (Pearson $r = 0.70$).

The Truth Was Never Gone: Perfect Aliasing in Compliant-Context Truth Probes

Sole author



- Established that compliant-context fitting labels cannot distinguish truth from prescribed action; complementary rival labels yield $\text{AUROC}_{\text{action}} = 1 - \text{AUROC}_{\text{truth}}$, verified across **751 cell-layer pairs**.
- Recovered truth with final-layer probes in reward-trained Gemma-2-9B: ally-fit AUROC **0.006** vs. mixed-fit **1.000** (three seeds, identical held-out activations); mixed fitting used more data and labelled rival contexts.

Internalizing the Halt Vector: What a Halt Target Changes in a Reasoning Model

First author

- Compared activation-target and text-loss effects in a 7B reasoning model with seeded replicates and random, reversed and covariance-direction controls; halt-specific mediation remains unresolved.
- On-axis training without text loss cut thinking by **34.3–45.6%** on MATH500, AIME and GPQA-Diamond; MATH500 any-position Math-Verify accuracy was non-inferior (3.85-point margin), with fewer correct boxed answers on MATH500/GPQA and no established cross-model transfer.

RL Training-Dynamics Interpretability

- Built a single-H200 GRPO/RLVR pipeline for a 7B model with dense LoRA checkpoints and held-out probing.
- Designed a pre-registered four-arm std×length normalization factorial for length-penalized GRPO, its baseline arm **bit-identical** to upstream TRL across **168 cases**; pipeline calibrated on cloud H200s, factorial unrun.

TECHNICAL SKILLS

Languages & Data: Python, C/C++, Java, CUDA, SQL, Bash, NumPy, pandas, scikit-learn, OpenCV, BigQuery.

ML & Training: PyTorch, Transformers, PEFT/LoRA, TRL (SFT, GRPO, DPO/SimPO), DDP/FSDP, mixed precision.

Performance & Retrieval: vLLM, TensorRT, torch.compile, torch.profiler, gradient checkpointing, FAISS.

Infrastructure: Docker, Kubernetes, Modal, GitHub Actions, pytest, ML regression CI, Weights & Biases.