Uday Tyagi

ut25@cornell.edu | (224) 551-1668 | github.com/heyuday | linkedin.com/in/heyuday

Cornell Master's in Computer Science student graduating December 2026 and available to start January 2027; ML/SWE engineer building agentic tooling in production; Focused on AI safety and alignment research: interpretability, steering, and multi-agent robotic planning. Research engineer on a NASA-funded multi-drone deconfliction system.

Education

Cornell University — Ithaca, NY
Master of Engineering in Computer Science, GPA 3.8/4.0 — *Jan. 2026 – Dec. 2026*

- Coursework: Advanced ML Systems, Neural Networks, Robot Learning, Deep Learning, Computer Vision, Algorithms, Operating Systems

University of Wisconsin–Madison — Madison, WI
B.S. in Computer Science, GPA 3.8/4.0 — *May 2023 – Dec. 2025*

- Completed the degree in two and a half years

Professional Experience

KLA — Remote
Software Engineer — *May 2025 – Present*

- Fine-tuned an LLM-powered AI agent for the DART web application, enabling engineers to interact with defect-inspection workflows in natural language, from image classification to event routing and triage
- Extended the DART agent with a custom MCP server exposing internal inspection and triage tools, giving the LLM structured, auditable tool-calling access instead of ad hoc API integrations
- Built visual regression and UI-performance monitoring pipelines for HPC automation workflows using Grafana Faro and custom metrics tooling, reducing debugging time in large-scale wafer inspection systems

Foresee Health — San Francisco, CA
Software Engineering Intern — *Sept. 2024 – Jan. 2025*

- Engineered an authentication system for the invoice application using AWS Amplify and Lambda
- Designed PostgreSQL schemas and synchronization patterns for a microservices architecture handling large customer datasets

Research

Research Engineer – NASA USRC (Project Orion) — Ithaca, NY
Cornell University, advised by Mehrnaz Sabet — *Jan. 2026 – Present*

- Trained a diffusion-based motion planner replacing ORCA for real-time multi-drone collision avoidance: DiT architecture generating velocity command sequences from ego and neighbor states, with DPM-Solver++ inference optimization and low-temperature reranking, reaching ∼80% mean scenario completion vs. ORCA's 59%
- Built the surrounding deconfliction stack: a ROS 2 planner plugin, scenario-based evaluation pipelines with behavioral analytics, and a ∼200k-trajectory training dataset across six collision scenarios

Researcher – LAISR (Long-Term AI Safety Research) — Ithaca, NY
Cornell University, advised by Prof. Lionel Levine — *Jan. 2026 – Present*

- Constitutional Drift: measuring whether an LLM's stated values remain stable when the model repeatedly rewrites its own constitution, tracking semantic drift, value collapse, and self-reinforcing edits across successive rounds of self-revision
- Built the experimental harness with two collaborators: multi-round self-editing pipelines across model families, embedding- and judge-based drift metrics, and behavioral evals testing whether edited constitutions change downstream model behavior

AI Safety Fellow – CAMBRIA Program — Cambridge, MA
Cambridge Boston Alignment Initiative — *May – June 2026*

- Worked through the ARENA mechanistic interpretability curriculum: linear probes, steering vectors, activation patching, sparse autoencoders, OthelloGPT, emergent misalignment, alignment faking, and LLM evals
- Original research using Natural Language Autoencoders (Fraser-Taliente et al., 2026): generated effective steering vectors for Qwen2.5-7B using natural language that shift model behavior across concepts and are geometrically concept-specific.

AI Safety Research Fellow Remote
AI Safety Camp *Dec. 2025 – April 2026*

- Evaluated hierarchical and parallel AI control protocols on the ControlArena benchmark, measuring safety-usefulness tradeoffs as the capability gap between trusted and untrusted models grows
- Extended ControlArena with red-team/blue-team protocols and Elo-based scoring to measure failure thresholds of AI oversight mechanisms under adversarial prompting

Researcher – SMARAG Ithaca, NY
Cornell University, Prof. Oliver Gao Lab *Jan. 2026 – Present*

- Built a multi-agent RAG system for automated ESG carbon reporting (LangChain orchestration, RL for factual grounding); designed an agent observability layer with audit-trail logging and behavioral guardrails to detect hallucinated citations in regulatory filings

Projects

NLA Steering: Text-Driven Behavioral Steering of LLMs | *Python, PyTorch, HuggingFace, Qwen2.5-7B*

- Used Natural Language Autoencoders (Fraser-Taliente et al., 2026) to generate steering vectors for Qwen2.5-7B from hand-written text descriptions alone — no training data, no activation extraction, no labeled examples; vectors reliably shift model behavior at inference time across multiple concepts (sycophancy, misalignment, persona)
- Running follow-on experiments ablating which parts of the text format carry the steering signal, and benchmarking the vectors as zero-shot behavioral probes against ground-truth activation-derived vectors

Real-Time Weather Stream Processor | *Apache Kafka, HDFS, PyArrow, Parquet, Docker, MySQL, gRPC, Python*

- Built a fault-tolerant streaming pipeline simulating nationwide weather stations, moving MySQL records into HDFS as partitioned Parquet with exactly-once semantics, checkpointing, and atomic writes
- Orchestrated a containerized multi-service stack (Kafka, MySQL, HDFS, gRPC) with Docker Compose, supporting horizontal scaling and stateful recovery after broker and node failures

Domain-Adaptive LLM Embedding Benchmark & Fine-Tuning | *Python, PyTorch, HuggingFace, sentence-transformers*

- Generated semantically adversarial "hard negative" passages via LLM prompting and used them as contrastive training data to fine-tune a retrieval model, improving retrieval accuracy across scientific and code domains
- Benchmarked 5+ embedding models and introduced a robustness metric quantifying how often each is fooled by near-miss passages

LLM Jailbreak Defense Evaluation | *Python, PyTorch, HuggingFace*

- Engineered jailbreak experiments combining GCG and AutoDAN automated attacks with layered defenses (self-reminders, hierarchical prompting, perplexity filters); evaluated guard-model architectures via Attack Success Rate metrics

Wisconsin School Analytics on GCP | *BigQuery, Dataform, GCS, SQL, Python*

- Built a cloud-native analytics pipeline integrating geospatial and public-school datasets across 400+ schools; automated Parquet ingestion via PyArrow with parameterized SQL on remote VMs

Leadership

Operations Lead – Cornell AI Alignment (CAIA) Ithaca, NY
Cornell University *Jan. 2026 – Present*

- Lead operations and semester planning for CAIA: coordinating fellowships, running the weekly technical paper reading group, and connecting students with AI safety research opportunities

Fundamentals Cohort Lead – Wisconsin AI Safety Initiative Madison, WI
University of Wisconsin–Madison *Sept. 2024 – Dec. 2025*

- Led a 10-student cohort through the AI Safety Fundamentals curriculum, facilitating weekly technical discussions on alignment, interpretability, and AI governance

Technical Skills

Languages: Python, Java, C/C++, JavaScript, TypeScript, SQL, CUDA
AI/ML: PyTorch, HuggingFace, mechanistic interpretability (probes, steering vectors, SAEs, activation patching), fine-tuning (LoRA/PEFT), evals, RAG, diffusion models, LangChain, MCP, inference optimization
Tools & Infra: Docker, Kubernetes, AWS (S3, Lambda), GCP, ROS 2, PostgreSQL, Kafka, Spark, React, Node.js