Vedaang Chopra

vedaangchopra@gatech.edu — +1 (404) 740-9905 — Atlanta, GA — LinkedIn — GitHub — Website

Research Engineer / Agentic AI Engineer — ML, LLM and Multimodal Systems

Large Language Models (LLMs) · AI Agents · LoRA Fine-Tuning · Evaluation · Vision-Language Models (VLMs) · ML Engineering

Work Experience

Georgia Institute of Technology — CORE Robotics Lab @ Siemens Atlanta, GA

Graduate Research Assistant — Advisor: ***Dr. Matthew Gombolay*** *Jan 2026 – Present*

- Built an AI agent (LangGraph, Python) that writes CAD code in CadQuery (the instructions that define a 3D part), runs it, compares the 3D shape with the correct one, and repairs its own mistakes in up to three tries. This is inference-time search and repair with a frozen large language model (LLM); no fine-tuning or reinforcement learning is used.
- Evaluated three systems on 200 expert and 200 beginner prompts: one-shot generation, self-debug, and agentic repair. Mean 3D shape overlap (IoU; 1.0 is a perfect match) rose from 0.82 to 0.92 (expert) and 0.57 to 0.70 (beginner) versus one-shot, and beat self-debug by about 0.05 (paired bootstrap, 95% intervals above zero). Repair uses the correct shape as guidance; a version without it is the next experiment.
- Built an evaluation harness with IoU, Chamfer and Hausdorff surface distances, and an optional vision-language model (VLM) judge used as an advisory signal, backed by a test suite of about 2,500 collected unit and integration tests.
- Built a process verifier that finds which step of a CAD program (feature tree, B-Rep geometry) caused a wrong shape: 60 of 60 injected faults detected and 53 of 60 correctly blamed first, on hand-built tests (not yet validated on real model errors).
- Reported negative results: richer feedback tied plain error messages (4 of 12 solved in every setup), and a pre-registered test of structural shape distance failed its bar (14.8% accuracy vs. 27.0% majority baseline), so it was not used in repair.

Fortinet Technologies Inc. (AIOps R&D) Bengaluru, India

Software Development Engineer I & II (ML/AI Track) *Feb 2021 – Jul 2025*

- Led an agentic retrieval-augmented generation (RAG) troubleshooting assistant from hackathon prototype (top-5 finalist, Fortinet Global Hackathon 2023) to production: the LLM plans multi-step tool calls over network telemetry (LangChain, structured function calling), verifies its hypotheses and reports a root cause. Mean time to resolve issues fell by ∼70% (self-reported).
- Re-architected the OpenSearch log-ingestion pipeline with async I/O in Go, raising throughput from under 50 to over 2,000 events per second (40×) with reliable, observable operation.
- Built and deployed 60+ machine-learning classifiers (scikit-learn, PyTorch) across 4 categories for 7-day network performance (SLA) forecasting, with automated evaluation and retraining.
- Built the unsupervised anomaly-detection method behind patent PCT/IN2022/058026 (Filed), cutting manual troubleshooting by ∼75% (self-reported); built DBSCAN anomaly detection on SD-WAN telemetry that reportedly prevented over 50% of potential outages.
- Optimized edge inference with ONNX Runtime and quantization (GPU to CPU), cutting latency by ∼40% (self-reported); integrated the data-science team's models into Python and Go backend services with CPU-only inference (FastAPI, Docker, PostgreSQL).

Education

Georgia Institute of Technology Atlanta, GA

M.S. in Computer Science *(Machine Learning),* ***GPA: 3.8/4.0*** *Aug 2024 – Dec 2026*

- **Coursework:** Deep Learning, Deep Reinforcement Learning, ML Security, Language Models and Language Agents, Large & Vision Language Models, Agentic AI, Systems for AI

Maharaja Surajmal Institute of Technology New Delhi, India

B.Tech. in Information Technology, CGPA: 8.8/10.0 *Aug 2016 – Aug 2020*

Selected Projects & Publications

CERBERUS — Vision-Language Model: Fine-Tuning and Cross-Modal Retrieval (GT research) Aug 2025 – Present

- Fine-tuned a 7B Qwen-2.5 model with LoRA supervised fine-tuning (SFT; rank 64 on attention layers, 40.4M trainable parameters) to answer questions about images. Training loss fell from 0.74 to 0.10 over 7,500 steps on a GPU cluster; a held-out evaluation is the next step.
- Trained a cross-modal alignment layer (PyTorch) over frozen CLIP and SBERT encoders with Matryoshka (nested) embeddings. On a PixMo validation set, the right caption ranked in the top 5 for 77.7% of images and the right image for 77.5% of captions at 4096 dimensions; at 128 dimensions, 62.9% and 64.0%.

ARTEMIS — Cost-Aware Routing Across Vision-Language Models (offline research) Aug 2025 – Present

- Built a leak-free benchmark of 67.7K questions and 338.6K responses from several VLMs (fixed images shared between train and test), and measured the best possible per-question routing: 65.4% correct versus 49.1% for the best single model (10,198 held-out questions).
- Trained a router (DistilBERT plus question metadata, PyTorch) to predict which model answers correctly. It did not beat a per-task lookup on the test set (50.9% vs. 56.4%), which we report as a negative result. vLLM serving configuration prepared (offline).

SHASTRA — Agent Trace Modeling & Constraint-Aware Planning (team research, GT) Jan 2026 – Present
- Built the tools that turn recorded agent benchmark runs (GAIA traces, Postgres) into step-by-step graphs: runs are split into blocks, an LLM labels what each block does, and an append-only label registry keeps names consistent. Wrote the analysis code for workflow reuse; the reuse measurement is not finished and retrieval is not built. Orchestrator and search were written by teammates.

ATHENA — Multi-Agent Idea-to-Storyboard Generation (team project, GT) 2025 – 2026
- Designed a 14-step LangGraph pipeline (multi-agent orchestration, typed Pydantic data records between steps) that turns an idea into a storyboard: scenes, shots, script and generated images; about 800 automated tests.
- Ran a 20-run stress test that exposed two bugs (scenes collapsing onto one beat, image-generation failures hidden as "done"); the fix is committed and being re-verified. Earlier version (2025): supervisor-worker agents with reflection.

Consensus Mutual-Fund Intelligence Platform — Data Engineering & Analytics 2026
- Built a PostgreSQL data platform (Python) that combines mutual-fund holdings and net-asset-value history from many fund houses, with data-quality quarantine rules and about 390 automated tests. A walk-forward backtest of the consensus signals showed negative returns, which we report as is.

Deep RL, Language Models and Security — Georgia Tech coursework & team projects 2024 – 2026
- Deep RL (team): trained a Proximal Policy Optimization (PPO) soccer agent with Ray RLlib and a 5-stage curriculum for 30M steps; it won 8 of 10 test games against the baseline.
- Language models: implemented byte-level BPE tokenization, a decoder-only Transformer with grouped-query attention, AdamW and KV-cache decoding from scratch (1.67× faster decoding at batch size 32).
- ML security: ran an embedding-poisoning backdoor (99.7% attack success, 93.3% clean accuracy) and a membership-inference attack on a fine-tuned language model (AUC 0.994).
- Graph neural networks (team of four, MalNet-Tiny): GraphSAGE reached 78.1% test accuracy (macro-F1 77.7%) and GAT 70.8% for malware classification.
- Smart-glasses sign reader (team of two, iOS): built OCR, database and live-video modules with a vision-language model; local OCR ran at 1.69 seconds per image on CPU.

Publications & IP
- **IEEE ICAIA 2026** (accepted; presenting Nov 16, 2026) — "Integrating Machine Learning and Memory Forensics for Enhancing Cybersecurity in IoT-Enabled Energy Systems" (with Sonika Malik): gradient-boosting ransomware classifier, 88.7% accuracy on 976 de-duplicated memory windows (exploratory).
- **Patent (Filed)** PCT/IN2022/058026 — Unsupervised distributional thresholding for wireless connectivity anomaly detection. **Ontology-based Text Classification**, CEUR Workshop Proceedings Vol. 2786 (2020).

Technical Skills

Agentic AI & LLM Systems: LangGraph, LangChain, AI agents, multi-agent orchestration, tool/function calling, agentic RAG, retrieval-augmented generation, planning and reflection, structured outputs (Pydantic), prompt engineering, LLM-as-a-judge, long-horizon task decomposition

Model Training & Fine-Tuning: LoRA / PEFT, supervised fine-tuning (SFT), PyTorch, Hugging Face Transformers, Qwen, DistilBERT, mixed precision, DDP-capable training code, experiment design and ablations

Post-Training Theory (coursework/self-study): RLHF, DPO, GRPO, RLVR, reward modeling

Reinforcement Learning (academic): PPO, DQN, policy gradients, curriculum learning, reward shaping, Ray RLlib

Evaluation & Benchmarking: model evaluation, benchmark design, execution-based evaluation, paired bootstrap, IoU, Chamfer/Hausdorff, Recall@k, A/B-style ablations

Multimodal & Vision-Language: VLMs, CLIP, SBERT, vision transformers, cross-modal retrieval, Matryoshka embeddings, OCR, computer vision, NLP

Inference, Data & Systems: ONNX Runtime, quantization, edge/CPU inference, model routing, vLLM (config-level, offline), FastAPI, OpenSearch/Elasticsearch, PostgreSQL, SQL, Redis, Docker, Linux, Git, Azure, W&B, TensorBoard

Security (coursework): adversarial attacks, data poisoning, backdoors, membership inference

Languages: Python, Go, C/C++, SQL, Bash