# Vedaang Chopra

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

**Research Engineer — ML/LLM Systems** — Model Training · Evaluation · Multimodal AI

## 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 that writes CAD code (the instructions that define a 3D part), runs it, compares the 3D shape with the correct one, and fixes its own mistakes in up to three tries. The language model is frozen; all gains come from the repair loop.
- Raised the average shape-overlap score (IoU; 1.0 is a perfect match) from 0.82 to 0.92 on 200 expert prompts and from 0.57 to 0.70 on 200 beginner prompts versus one-shot generation, and beat a self-debug baseline by about 0.05 (95% intervals above zero).
- Built the scoring tools (IoU, Chamfer, Hausdorff) and a checker that finds which program step caused a wrong shape (60 of 60 injected faults on hand-built tests). Repair still uses the correct shape as guidance; removing that is the next experiment.

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

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

- Rebuilt the pipeline that collects network-device logs (OpenSearch, async code in Go), raising its speed from under 50 to over 2,000 events per second (40×).
- Took an AI troubleshooting assistant from a hackathon prototype (top-5 finalist, 2023) to production: the model plans steps, calls tools on network data and checks its own guesses, cutting mean time to resolve issues by ∼70% (self-reported).
- Built the anomaly-detection method behind patent PCT/IN2022/058026 (Filed), cutting manual troubleshooting by ∼75%, and cut edge-model latency by ∼40% with ONNX Runtime (both self-reported).

## Education

**Georgia Institute of Technology** Atlanta, GA

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

**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** — Teaching a Language Model to Understand Images (GT research) Aug 2025 – Present

- Fine-tuned a 7B Qwen-2.5 model with LoRA (rank 64; trains a small add-on, not the whole model) to answer questions about images. Training loss fell from 0.74 to 0.10 over 7,500 steps; a held-out test is the next step.
- Trained a layer that matches images to captions on frozen CLIP and SBERT encoders. On a PixMo validation set (4096-number embeddings), the right caption ranked in the top 5 for 77.7% of images and the right image for 77.5% of captions; at 128 numbers, 62.9% and 64.0%.

**ARTEMIS** — Choosing the Cheapest Vision-Language Model per Question (offline research) Aug 2025 – Present

- Built a leak-free benchmark of 67.7K questions and 338.6K model answers (fixed shared images across splits) and measured the best possible routing: 65.4% correct versus 49.1% for the best single model (10,198 held-out questions).
- Trained a router (DistilBERT plus question metadata) 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.

**Coursework** — Deep RL and Language Models (Georgia Tech) 2026

- Team project: trained a PPO soccer agent (Ray RLlib, 5-stage curriculum, 30M steps) that won 8 of 10 test games. Built a small Transformer language model from scratch, with fast decoding (KV-cache, 1.67× faster).

**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). **Patent (Filed)** PCT/IN2022/058026.

## Technical Skills

**Model Training:** LoRA fine-tuning (PEFT), PyTorch, Hugging Face Transformers, PPO (coursework), DDP-capable training code, ablations

**Evaluation & Multimodal:** benchmarking, IoU, Chamfer/Hausdorff, Recall@k, paired bootstrap, VLM evaluation, CLIP, SBERT

**Agents & Inference:** LangGraph, LangChain, tool calling, agentic RAG, LLM-as-a-judge, ONNX Runtime, vLLM (config-level, offline)

**Programming:** Python, Go, SQL/PostgreSQL, OpenSearch, FastAPI, Docker, Linux, Git, scikit-learn