

LEONARDO SCHAUB

(512) 586-6398 · leoschaub28@gmail.com · github.com/lmfor

Education

Texas A&M University

B.S. in Computer Science, Minors in Math and Japanese

May 2028

College Station, Texas

Experience

Qualcomm Inc.

May 2026 – August 2026

Software Engineering Intern | NOVA (Next-Generation Observability Virtual Agent)

San Diego, California

- Migrated NOVA from a monolithic agent to a scalable **multi-agent Swarm Architecture**, achieving **86% lower token usage** and **72% faster response times**; prototyped Supervisor and Supervisor/Swarm hybrid designs.
- Designed and optimized **concurrent tool-result compression** and de-duplication in **LangGraph**, cutting NOVA response time by **~46%** while reducing token overhead.
- Built an **agent-regression benchmarking framework** using golden test cases, tool-call analysis, and **LLM-as-a-judge** rubrics to compare accuracy, latency, and token efficiency across NOVA releases.

Advantest America Inc.

May 2025 – August 2025

Machine Learning Engineering Intern

Austin, Texas

- Reduced SmartTest 7/8 graph generation time by **80%+** by building an **NLP-driven interface** that translated natural language requests into automated **SQL** queries and visualizations.
- Designed **agent orchestration infrastructure** using **MCP servers** to safely execute external tools and enforce structured workflows across LLM systems.
- Improved LLM-generated analysis pipelines by implementing post-query **anomaly detection** and validation for agent-generated outputs.
- Automated updates to the Test Documentation Center (TDC) by generating relevance reports that identified and submitted new documents for **RAG** retraining, reducing manual document review time by **60%+**.

Research

Texas A&M University

March 2026 – Present

Undergraduate Machine Learning Researcher

College Station, Texas

- Researching **LLM inference and architecture optimization**, experimenting with **speculative decoding**, **quantization**, **KV-cache optimization**, and **batching** to reduce inference latency, memory usage, and computational cost.
- Investigating architectural modifications including **Mixture-of-Experts (MoE)**, dense-to-MoE transformations, and **multi-agent architectures** to study tradeoffs between model capacity, routing, parallelism, and inference efficiency.
- Experimenting with **abliteration** and representation-level model modifications to analyze how targeted weight and activation interventions affect model behavior and downstream performance.
- Building a reproducible **PyTorch inference harness** to benchmark experiments using **TTFT**, **inter-token latency**, **tokens/sec**, **peak memory**, **token usage**, and **output quality**.

Projects

LoL ML Pipeline | *Python, TensorFlow, Pandas, Docker, AWS, Snowflake*

2026

- Built an **end-to-end ML pipeline** predicting professional League of Legends match outcomes from **pre-game signals only** (team and player identity, draft picks and bans, side selection, patch context), excluding post-game statistics to **eliminate label leakage**.
- Designed **deterministic feature engineering** over versioned raw match data with **immutable dataset snapshots**, making every training run reproducible from a resolvable dataset identifier.
- Architected **experiment tracking**, **model artifact versioning**, and **drift monitoring** hooks to support scheduled retraining, prioritizing system design correctness over short-term model accuracy.

Technical Skills

Languages: Python, C++, Java, Haskell, SQL

ML / AI: TensorFlow, PyTorch, LangGraph, MCP, RAG, Transformers, LLM Inference Optimization

Data / Observability: NumPy, Pandas, PostgreSQL, Splunk, Datadog, Langfuse, Langsmith

Cloud & DevOps: AWS (S3, Batch, ECR), Docker, Git, Linux, CI/CD Pipelines

Leadership / Extracurricular

Texas A&M College of Engineering Student Ambassador

Fall 2025 - Present

- Lead engineering tours for prospective students, families, and VIPs, engaging approx. 400 participants weekly.

Honors: Dean's Honor Roll (Spring 2025, Fall 2025); Deans' Excellence Award (Spring 2026)