suramyaangdembay.github.io
Suramya.Angdembay@usm.edu
GitHub

Suramya Raj Angdembay

Research Interests

Mechanistic Interpretability, Chain-of-Thought Faithfulness, Trustworthy AI, Sparse Autoencoders, Causal Interventions, Validity of LLM Benchmarks, Forecast Error Analysis

Education

University of Southern Mississippi, Hattiesburg, MS **Aug. 2023 – May 2027 (expected)**
B.S. Computer Science and B.S. Mathematics, Honors College GPA 3.71/4.0
– Honors Keystone Scholar and merit-based full-tuition scholarship recipient; President's List (two terms) and Dean's List, 2023–2025.

Preprints and Manuscripts

[1] **S. R. Angdembay**, D. Aryal, and N. Rahimi. *Two Regimes of Chain-of-Thought Unfaithfulness: Metric-Based Detection Fails Where Models Are Wrong.* arXiv preprint; under review at ACL Rolling Review, 2026. [arXiv] [code]

[2] **S. R. Angdembay** and H. Tian. *What Do Benign-Trained Session Language Models Actually Detect? A Mechanistic Audit on Insider-Threat Benchmarks.* Manuscript in preparation, 2026.

Research Experience

Chain-of-Thought Faithfulness in Language Models [1] **May 2026 – Present**
Advisor: Dr. Nick Rahimi · UR2PhD program (CRA-WP)
– Audited behavioral detectors of unfaithful chain-of-thought on FaithCoT-Bench by answer correctness: where models are wrong, and most unfaithfulness occurs, no tested signal was above chance.
– Showed answer incorrectness alone beats every purpose-built detector and the step-removal metric anti-correlates with human labels; linear probes recover signal behavioral detectors miss.

Interpretable Insider-Threat Detection [2] **Aug. 2024 – Present**
Advisor: Dr. Haiyan Tian
– Designed and ran a comparison of user-activity representations (PCA to sparse features of a QLoRA-adapted LLM) on the CERT benchmarks; basis of the Honors Thesis and two talks.
– Built a mechanistic audit of what a benign-trained language model detects: sparse-autoencoder decomposition, causal interventions against matched controls, token-level attribution.

ML Error Analysis of NOAA Ocean Forecasts **Jan. 2026 – Present**
Advisor: Dr. Gregg Jacobs · Institute for Advanced Analytics and Society
– Showed RTOFS's hurricane-heat-potential error against Argo is largely systematic (too low in the tropics, shared with GOFS); a correction learned from Argo removes about a third of it on forward tests.
– Found random-month holdouts overstate skill (errors persist for weeks to months), so tests use forward-in-time blocks; noise-column and shuffle tests showed a quarter of inputs carry no signal.
– Showed RTOFS's own 7-day change is the most useful input and anti-correlates with its error (RTOFS overshoots), and its detail below about 50 km is mostly unrealistic, with no sign of misplaced fronts.

Presentations and Talks

- **S. R. Angdembay** and H. Tian. *Principal Component Analysis for Data Compression, Noise Reduction, and Anomaly Detection: A Case Study on Insider Threat Detection.* AMS Spring Southeastern Sectional Meeting (undergraduate research session), Clemson University, Mar. 2025.
- **S. R. Angdembay** and H. Tian. *Continuous Time Markov Chains for Insider Threat Behavior Modeling.* 33rd USA/USM/SELU Research Mini-Conference, Apr. 2026.
- Lightning talk, SC25 Student HQ (ACM SIGHPC Immersion Program), St. Louis, Nov. 2025.

Honors, Fellowships, and Awards

- **Break Through Tech AI Fellow** (Cornell Tech), 2026 – present.
- **ACM SIGHPC Immersion Program, SC25** — 1 of 15 students selected nationally, 2025.
- **Honors Keystone Scholar Award** ($2,000 research grant), USM, 2025 – 2027 — funds the Senior Honors Thesis.
- **Wright W. & Annie Rea Cross Mathematics Research Scholarship** ($2,000), USM, 2025.
- **eBay University Machine Learning Challenge** — 23rd on the public leaderboard, 2025.
- **COMAP Mathematical Contest in Modeling**, Successful Participant, 2025.
- Summer programs: CompuCell3D Biological Modeling (Indiana University); WISER & Womanium Quantum, 2025.

Additional Experience

AQUAVIEW Ocean Data Platform **Jan. 2026 – Present**

Supervisor: Dr. Josh Hill · Institute for Advanced Analytics and Society

- Built the data-ingestion pipeline and the review-and-publish controls for a NOAA-facing ocean data platform; shipped to production.

Mathematical Modeling and Optimization Intern **Jan. – May 2026**

Optimal Answers LLC

- Encoded legislative rules and scoring metrics for Gulf Coast Restoration Fund projects into databases for optimization-based allocation.

Weather API Developer **Jul. – Dec. 2025**

AireChemistry

- Designed a FastAPI service serving gridded temperature, wind, and $PM_{2.5}$ fields for South Africa.

Technical Skills

- **Research methods:** sparse autoencoders, causal patching and ablation, linear probing, LLM-as-judge evaluation, bootstrap and permutation inference.
- **ML:** PyTorch, Hugging Face Transformers, PEFT/QLoRA, scikit-learn, Qiskit. **Languages:** Python, C/C++, R, MATLAB, JavaScript.
- **Systems:** Slurm on Purdue Anvil (A100/H100) and USM Magnolia, Linux, Git/GitHub, FastAPI, LaTeX.

Service

- Volunteer Teaching Assistant, Jubilant Elite Academy, Nepal (2022); Field Volunteer, Human Rights Forum Nepal (2022).