

Shubham Gaur

Santa Clara, California, USA

✉ gshubhaur@gmail.com

🐙 GitHub

🌐 Portfolio

🌐 LinkedIn

🎓 Google Scholar

Summary

First-year Ph.D. student at UC Santa Cruz (advisor: Dr. Ian Lane) researching LLM post-training, factuality, and safety-focused evaluation of LLMs and agents. Work spans agent benchmarks, reward hacking in agent evaluation, and user dependence on supportive LLMs, with a co-first-author ICLR 2027 submission and 1st place at UC Berkeley AgentBeats 2026. Brings 6 years of production ML engineering at 100M+ user scale.

Education

University of California, Santa Cruz <i>Ph.D., Computer Science (expected Dec 2029) • Advisor: Dr. Ian Lane</i>	Sept 2026 – Present Santa Cruz, USA
University of California, Santa Cruz <i>M.S., Natural Language Processing (AI) – 4.0/4.0 • Advisor: Dr. Chenguang Wang</i>	Sept 2024 – March 2026 Santa Clara, USA
SRM Institute of Science and Technology <i>B.S., Computer Science – 3.8/4.0 • Advisor: Dr. A. Shanthini</i>	July 2015 – May 2019 Chennai, India

Publications & Preprints

*Equal contribution (co-first author).

- [1] Shukla, S.*, Kim, J.*, Gaur, S.*, Roshanaei, M., & Seif El-Nasr, M. (2026). RELATE: An Evaluation Framework for Measuring Relational Orientation of Large Language Models. *Under review at ICLR 2027*. arXiv:2610.09569.
- [2] Zheng, S., Di, Z., Liu, Y., ..., Gaur, S., Ye, Z., Wang, H., Li, X., Song, D., & Hauser, C. (2026). BenchShield: Formal Model-Backed Instrumentation for Reward Integrity in LLM-Agent Evaluation Infrastructure. *Preprint*. arXiv:2609.11028.
- [3] Yagnik, S., Gaur, S., Thakur, S., Jain, V., Chadha, A., & Das, A. (2026). MAAT: Multi-phase Adapter-Aware Targeted Unlearning. *Preprint*. arXiv:2605.30514.
- [4] Gaur, S., & Lane, I. (2026). Signal-Driven Observation for Long-Horizon Web Agents. *ICML 2026 Workshop on Failure Modes in Agentic AI (FAGEN), Position Paper*. arXiv:2606.06708.
- [5] Crispino, N., Gaur, S., Yang, X., Yu, A., Ercevik, B., Sapugay, C., Potter, Y., Song, D., & Wang, C. (2026). SkillOptimizer: Agent Skill Optimization Through Subskills Without Task Supervision. *ICML 2026 AIWILD: Safety, Security, and Beyond Workshop*.
- [6] Tu, J., Gaur, S., Murtinty, R., Wang, Z., Shi, T., Song, D., & Wang, C. (2026). FAULTLOC: Evaluating AI Coding Agents for Fault Localization from Crash to Cause. *ICML 2026 AIWILD: Safety, Security, and Beyond Workshop*.
- [7] Clymo, J., Zernik, A., & Gaur, S. (2025). UCSC NLP T6 at SemEval-2025 Task 1: Leveraging LLMs and VLMs for Idiomatic Understanding. *Proceedings of SemEval-2025 (ACL Workshop)*, pages 2103–2115.
- [8] Kilaru, S., Gaur, S., & Rout, S. (2024). Advancing Web-Based Visual Question Answering with Efficient Image-Text Alignment. *Recent Advancements in Artificial Intelligence: Proceedings of ICRAAI 2023*, pp. 343–359. Springer.
- [9] Naulakha, R., Gaur, S., Lodha, D., Tulsyan, M., & Kotecha, U. (2021). Extraction of Cumulative Blobs Using Dynamic Gestures. *International Journal of Scientific Research (IJSR)*, 10(9).

Research Experience

Graduate Researcher

Apr 2026 – Present

University of California, Santa Cruz • Advisor: Dr. Ian Lane

Santa Cruz, USA

- Investigating **long-tail factual recall** in LLMs via reasoning and post-training.
- Defined *observation over-ingestion* as a root cause of context rot and goal drift in long-horizon web agents; proposed **Signal-Driven Observation**, re-reading the DOM only on browser-event signals (first author, ICML 2026 FAGEN position paper). [\[Paper\]](#)

Graduate Research Assistant

Oct 2025 – Aug 2026

Dr. Chenguang Wang's Research Lab • with UC Berkeley RDI (Prof. Dawn Song)

Santa Clara, USA

- Designed **FAULTLOC**, a 900-CVE benchmark evaluating AI coding agents on fault localization; ran ablations across frontier models, showing frontier agents recover the true root cause in only 40% of cases. [\[View Code\]](#)
- Built **RCABench**, benchmark infrastructure for multi-hop causal reasoning and root-cause localization over large codebases; surfaced systematic multi-agent coordination failures under realistic security conditions. [\[View Code\]](#)
- Developed **SkillOptimizer**, improving LLM agent skills without task supervision via subskill DAG decomposition and coverage-driven optimization; raised task pass rates across 192 skills and three agent families.
- Contributed to **MassGen**, an open-source multi-agent LLM framework for coordinating heterogeneous models on complex, open-ended tasks. [\[View Code\]](#)

Open-Source Contributor

Jul 2026 – Present

BenchFlow

Remote

- Adjudicated reward-hacking trajectories for 35 tasks against a structured rubric for **BenchShield**, contributing to a 456-trajectory human-adjudicated corpus drawn from 31,000+ agent runs (co-author). [\[Paper\]](#)
- Contributing to **FrontierPhysics**, a benchmark of specialist-level physics research tasks: surveyed physics-evaluation literature, built a platform connecting the project with Nobel laureate physicists, and fixed bugs in BenchFlow's evaluation infrastructure. [\[Project\]](#)

Cross-Institution Research Collaborations

2026

UCSC GUII Lab, Stanford, BITS Pilani

Remote

- **RELATE** (with UCSC GUII Lab and researchers at Stanford): co-developed a framework measuring whether supportive LLMs foster reliance on the AI or encourage human connection; across 1,596 dialogues and 7 LLMs, AI-centered language rose from 40.5% to 72.9% over six turns (co-first author, under review at ICLR 2027). [\[Paper\]](#)
- **MAAT** (with researchers at Stanford and BITS Pilani): co-developed a three-phase LoRA-adaptor unlearning method and **5WBench**, a benchmark for causal-knowledge unlearning; MAAT retained 36.4 points more knowledge than the strongest baseline at equal forgetting. [\[Paper\]](#)

Research Apprentice

Apr 2025 – Dec 2025

Samsung Research America • Advisor: Dr. Beth Ann Hockey

Santa Clara, USA

- Built a **novel data collection pipeline** for capturing high-quality human interaction trajectories on ServiceNow using the WorkArena benchmark.
- Designed end-to-end preprocessing and post-processing of JavaScript interaction events to BrowserGym compatible actions. [\[View Code\]](#)
- Collected **50+ human trajectories** on WorkArena L1 tasks, showing a 30s reduction in agent (GPT-5) execution time via in-context learning.

Graduate Research Assistant

Apr 2025 – Jun 2025

Dr. Jeffrey Flanigan (NLP 244 Coursework)

Santa Clara, USA

- Benchmarked **diffusion language models (LLaDA)** against autoregressive LLMs (LLaMA-3, SmolLM) for faithfulness on summarization datasets. [\[View Code\]](#)
- Analyzed robustness under zero-shot, chain-of-thought, and adversarial prompting settings.

Industry Experience

Applied Research Intern

Jun 2025 – Aug 2025

Nokia Bell Labs

Naperville, USA

- Designed a **multimodal retrieval pipeline** for enterprise knowledge search with vector databases (FAISS, Chroma); evaluated retrieval latency and ranking quality with custom metrics and A/B tests.
- Built transformer-based time-series forecasting and **graph neural network** anomaly detection; optimized serving via GPU acceleration and ONNX conversion.

Founding Lead Machine Learning Engineer

Nov 2021 – Aug 2024

Adani Group

Gurgaon, India

- Architected ML infrastructure for a **100M+ user** personalization system; ran controlled A/B experiments across 50+ model variants, improving engagement by 35%.
- Built an end-to-end PyTorch training and deployment pipeline with data/model drift monitoring and automated retraining, cutting model refresh time by 60%.
- Part of the team that won the **MongoDB APAC Innovation Award 2023** (From Batch to Real-Time) for re-architecting the AdaniOne SuperApp data platform from batch to real-time event streaming.

Machine Learning Engineer

Jul 2019 – Oct 2021

BlackRock

Mumbai, India

- Built deep neural networks for fixed-income and equity portfolio optimization, cutting inference latency 60% via quantization and pruning.
- Developed risk-scoring ensembles with SHAP explainability and statistically significance-tested model comparisons.

Teaching Experience (Teaching Assistantships)

TIM 50: Business Information Systems

Fall 2026

ARTG 91: Intro to Game Art Production

Winter 2026

THEA 80C: Monsters

Fall 2025

TIM 172B: Introduction to Technology Management II

Winter 2025

TIM 172A: Introduction to Technology Management I

Fall 2024

Honors & Awards

- Recipient, **UC Santa Cruz Regents Award** and **International Doctoral Recruitment Fellowship** (2026).
- Winner, **NVIDIA GTC 2026 PyTorch Helion Hackathon** (LLM-guided Triton kernel optimization).
- **1st Place**, Cybersecurity Agent track, UC Berkeley AgentBeats Competition 2026.
- **BlueDot Impact AI Safety Fellowship** (2026), a selective cohort-based AI alignment program.
- Recipient, **Thinking Machines Lab Tinker Research Grant** (2025), with Dr. Chenguang Wang.
- NVIDIA Certified Associate & Professional: Generative AI LLMs.

Service

- **Peer Reviewer**: ICLR 2026, ACL 2025.
- **Open source**: MassGen (multi-agent LLM framework); BenchFlow evaluation infrastructure.

Poster Presentations

What Web Agents Need Isn't Data – It's High Quality Data

Capstone Presentation, UC Santa Cruz

2025

Advancing Web-Based Visual Question Answering with Efficient Image-Text Alignment

SciML Symposium, Georgia Tech

2024

Technical Skills

- **Languages:** Python, C++, SQL, Java
- **ML Frameworks:** PyTorch, JAX, Hugging Face Transformers, TensorFlow, scikit-learn
- **LLM Training & Post-Training:** Supervised Fine-Tuning (SFT), Reinforcement Learning from Human Feedback (RLHF), Direct Preference Optimization (DPO), LoRA/PEFT, Machine Unlearning, Knowledge Distillation
- **LLM Agents & Evaluation:** Agent Benchmarks, LLM-as-a-Judge, Reward Hacking Analysis, Multi-Agent Systems, Human Evaluation, Ablation Studies, Statistical Significance Testing
- **Systems & Infrastructure:** Distributed Training, CUDA, Triton, ONNX, Ray, Docker, Kubernetes, Spark, AWS SageMaker, GCP Vertex AI