

Marilyn Rego

mrego@umich.edu | <https://marilynrego.github.io> | www.linkedin.com/in/marilyn-rego

Ann Arbor, MI | (765) -476-6131

RESEARCH INTERESTS

• Distributed AI Systems • Heterogeneous Computing • ML Systems & Infrastructure • LLM Inference Systems • Hardware–Software Co-Design

EDUCATION

University of Michigan

Ph.D. in Computer Science & Engineering

Advisor: Muhammad Shahbaz

Ann Arbor, MI

Aug 2025 - May 2030

Purdue University

Bachelor of Science, Computer Science GPA 3.9/4.0

Concentration in Machine Intelligence | Minors: Mathematics, Accounting, Management

West Lafayette, IN

Aug 2021 - May 2025

TECHNICAL SKILLS

Languages: C, Java, Python

Frameworks & Tools: PyTorch, vLLM, Triton (kernel development), CUPTI (GPU profiling), Nsight Systems

Systems: Distributed LLM inference, multi-GPU profiling, heterogeneous CPU–GPU performance analysis

ML: Large language models, quantization, activation compression, adapter-based methods, model serving/inference optimization

RESEARCH AND PROFESSIONAL EXPERIENCE

University of Michigan, SRC Research Scholars Program

Graduate Student Research Assistant

Ann Arbor, MI

August 2025 - Present

- Researching communication-efficient systems for cross-region LLM inference through hardware–software co-design across geographically distributed data centers.
- Designing learned activation-compression techniques and adapter-based methods to reduce cross-region data movement while preserving model quality.
- Evaluating partitioning strategies for prefill, decode, and agentic inference workloads across latency, throughput, data movement, and model quality under heterogeneous compute and network constraints.

Adobe

Research Scientist/Engineer Intern

Seattle, WA

May 2026 - August 2026

- Designed and implemented CPU–GPU profiling infrastructure using CUPTI, including GPU kernel attribution, clock calibration, and support for heterogeneous GPU configurations.
- Developed a Python profiling plugin with automatic instrumentation and decorator-based APIs, backed by a native C++ profiling engine.
- Extended profiling to multi-process LLM inference workloads (vLLM, tensor parallelism, pipeline parallelism) and multi-node distributed systems, with automated trace collection and cluster teardown.
- Validated framework correctness and performance through structural trace verification and end-to-end testing on PyTorch and vLLM workloads.

Purdue University, Department of Computer Science, NextGArchLab

Undergraduate Student Research Assistant

West Lafayette, IN

Jan 2024 - August 2025

- Implemented and benchmarked classification models (one-class SVMs, random forests, isolation forests) for anomaly detection at the application payload level, establishing baseline metrics for accuracy, computational cost, and scalability.
- Analyzed the impact of packet-level, flow-level, and connection-level payload features on malicious traffic classification, evaluating trade-offs between prediction accuracy and computational overhead.
- Explored transformer-based models to automate feature selection for application payloads, comparing

performance against traditional methods integrated with SNORT.

Purdue University, School of Electrical and Computer Engineering

West Lafayette, IN

Summer Undergraduate Research Fellowship

May 2024 - August 2024

- Evaluated LLM performance in generating separation logic specifications for static program verification using VeriFast, progressing from baseline prompt engineering to chain-of-thought, prompt chaining, and few-shot prompting.
- Developed tailored input datasets and categorized VeriFast benchmarks to enable systematic model evaluation across prompting strategies.
- Designed a qualitative analysis framework to measure output correctness, syntactic validity, and model failure modes for LLM-generated specifications.

Purdue University, Department of Chemistry

West Lafayette, IN

Undergraduate Research Assistant

May 2023 - August 2023

- Implemented a Python deep-learning pipeline for predicting 3-dimensional molecular structures and properties from user-input parameters in the project – “Deep Learning for Drug Discovery”.
- Applied machine-learning methods to model crystal growth mechanisms and improve prediction accuracy.

PUBLICATIONS

Conference Papers

ISSRE. Wen Fan, Minh Tran, Sanya Dod, Xin Hu, **Marilyn Rego**, Danning Xie, Jenna DiVincenzo, and Lin Tan. *An Empirical Study of LLM-Generated Specifications for VeriFast*. September 2026.

FORGE. **Marilyn Rego***, Wen Fan*, Xin Hu, Sanya Dod, Zhaorui Ni, Danning Xie, Jenna DiVincenzo, and Lin Tan. *Evaluating the Ability of GPT-4o to Generate Verifiable Specifications in VeriFast*. April 2025. (*co-primary author)

Workshop Papers

NAIC Workshop @ SIGCOMM. Regan McDonald, **Marilyn Rego**, Ertza Warraich, Annus Zulfiqar, and Muhammad Shahbaz. *Towards Network-Efficient Cross-Regional Inference via Learned Activation Compression*. August 2026.

SELECTED PRESENTATIONS

mrLLM: Fast Multi-Region LLM Inference Using Learned Adapters. Poster, USENIX NSDI, May 2026.

SpliDT: Partitioned Decision Trees for Scalable Stateful ML Inference at Line Rate. Talk and poster, SRC TECHCON, Sept. 2025.

AI for Software Verification. Talk, Purdue Undergraduate Research Symposium, July 2024.

AWARDS AND HONORS

Adobe Research Code Quality Jam – Robustness Category Winner, 2026

USENIX NSDI Student Grant Awards, 2025 and 2026

Elected Member, Tau Beta Pi (TBP) Engineering Honor Society, 2026

First-Year CSE PhD Rackham Fellowship, 2025

Purdue College of Science Dean’s List and Semester Honors 2021-2025

PROFESSIONAL EXPERIENCE

Judge, University of Michigan Undergraduate Research Symposium

2026

MENTORING EXPERIENCE

- Maxwell Kumbong (SWE at Duolingo) - Distributed AI Systems, 2025-Present
- Yuchen Wang (UMich undergrad) - Distributed AI Systems, 2026-Present
- Regan McDonald (UMich undergrad) - LLM Activation Compression, 2025-2026
- Peer Mentor, CSEG Buddy Program, University of Michigan, 2026-Present

TEACHING EXPERIENCE

Undergraduate Teaching Assistant, CS 182: Foundations of Computer Science

Spring 2023