

Sanjay Sriram

Sanjaysriram44@gmail.com | +1 (765) 532-8987 | sanjaysriram44.com | linkedin.com/in/sanjaysriram44 | github.com/sanjaysriram44

EDUCATION

Purdue University

Bachelor of Science in Computer Engineering

Expected May 2028

West Lafayette, IN

- GPA: 3.8/4.0; Dean's List (2x), Semester Honors (2x); Coursework: Data Structures (C++), Advanced C, Systems

EXPERIENCE

Research Engineering Intern

Rox (Sequoia/General Catalyst, \$1.2B valuation)

May 2026 – Present

San Francisco, CA

- Built lifecycle hooks, heartbeats, distributed tracing, and observability into the platform orchestrating Rox's production agent swarms (Temporal, AWS, Docker, Kubernetes).
- Rebuilt the agentic retrieval layer over the production CRM, turning a relational schema into a knowledge graph: authored an OWL ontology and OBDA mappings with Ontop rewriting agent-issued SPARQL into SQL over Snowflake, so agents traverse stored edges instead of inferring relationships foreign keys never captured.
- Curated a benchmark of natural-language CRM questions and ran both retrieval layers behind an identical agent loop over MCP servers, across 8 model families from 27B to 1.6T parameters: graph retrieval raised agent accuracy from **28.8% to 97.3%** while cutting **tokens 37%, tool calls 42%, and cost 39%**; a 27B model on the graph beat a 1.6T model on tables.

Applied ML Engineer Intern

Ernst & Young (EY)

Jun 2025 – Aug 2025

Bangalore, India

- Shipped a containerized service turning plain-English system descriptions into enterprise architecture diagrams, collapsing a manual design task from hours to under 30 seconds.
- Built an extraction agent pulling GST and tariff schedules off Indian government portals and returning them as structured records, eliminating manual lookup for the tax practice.
- Rebuilt the search layer behind Business Blueprint (BBP) documents, pairing keyword and semantic retrieval over prior engagements so drafts assemble in under an hour instead of a week.

RESEARCH

Undergraduate Researcher

Human-Centered Software Systems Lab (Prof. Tianyi Zhang)

Aug 2026 – Present

West Lafayette, IN

- Building the AI software stack vulnerabilities graph in SecureChain, an NSF-funded knowledge graph selected as 1 of 10 datasets for NSF's **National AI Research Resource (NAIRR) Pilot**, modeling packages, model weights, LoRA adapters, training datasets, RAG corpora, vector stores, prompts, tool definitions, MCP servers, and agents as typed components.
- Built the pipeline populating it from public security literature (GitHub advisories, huntr bounties, vendor threat research), routing structured advisories to a deterministic parser and prose to schema-constrained LLM extraction, with source-span verification and entity resolution before any write.

Undergraduate Researcher

Duality Lab (Prof. James C. Davis)

Jan 2026 – May 2026

West Lafayette, IN

- Led framework evaluation for SysLLMatic, an RL-based post-training framework that aligns code models toward efficiency via DPO over 88K+ C++ preference pairs, optimizing runtime, memory, CPU cycles, throughput, and energy simultaneously rather than one metric at a time.
- Identified and curated benchmark datasets spanning program classes, workloads, and computing metrics to test whether the aligned policy generalized past its original single-dataset scope.
- Aligned policy delivered up to **39.5% CPU-cycle reduction** and 1.15× speedups on held-out C++ programs while preserving semantics.

PROJECTS

PebbleGPT: Small Language Model Pretraining for \$60 | PyTorch, Transformers, CUDA

Aug 2026

- Pretrained a 320M-parameter transformer on 10.4B tokens using GQA, SwiGLU, RoPE, and a WSD learning rate schedule, for \$50 on a rented H100.
- Diagnosed underperformance on PIQA and ARC-easy and ran a targeted \$1 mid-training pass: 300M tokens mixing 40% replay of the pretraining data with Cosmopedia's wikihow/stories splits (physical reasoning) and openstax/khanacademy splits (science reasoning), raising PIQA 55.5→67.7 and ARC-easy 37.7→56.4.
- Supervised fine-tuned on SmolTalk; total cost \$60. Beat GPT-2 (124M) on ARC-easy/ARC-challenge at a comparable token budget, and scored within a few points of Pythia-410M and SmolLM2-135M despite those models training on 30–200× more tokens.

TECHNICAL SKILLS

Languages: Python, C, C++, CUDA, SQL, SPARQL, JavaScript/TypeScript, MATLAB

ML & AI: PyTorch, Transformers, Triton, DPO/RLHF, QLoRA/PEFT, LangGraph, MCP

Infrastructure: Docker, Kubernetes, AWS, Snowflake, Temporal, FastAPI, Postgres/pgvector, Linux, Nsight