Satyam Goyal

sagoyal@umich.edu | satgoy152.github.io | linkedin.com/in/satyam-goyal-me | github.com/Satgoy152

Education

University of Michigan, Ann Arbor — Ann Arbor, MI
M.S. in Computer Science; GPA: 3.95/4.0 — *Expected May 2026*

- Coursework: NLP, DS&A, Computer Vision, Databases, Parallel Programming w/ GPUs
- Confrences: AAAI 24', Y Combinator AI Startup School 25'

Experience

IBM Research — May 2025 – Aug. 2025
Generative AI and Cloud Intern — *New York, NY*

- **Decreasing inference times for LLMs** hosted on distributed systems by modeling LLM serving.
- Developed an **open source vLLM simulator** that predicts industry SLO metrics with 95% accuracy.
- Used **Discrete Event Simulation** in Go and Bayesian optimizers to **decrease runtimes by 500x.**
- Published guides explaining vLLM internals to make accelerated inference serving accessible.

UofM CSE — Aug. 2024 – Present
UMich Research Assistant - First Author — *Ann Arbor, MI*

- Developing a novel benchmark to evaluate context storing AI-Agents on fitting to user preferences over multiple conversations.
- Improved **LLM** accuracy by 18% on game-theory problems using novel data representation techniques.

Optiwise.ai Inc. — May 2024 – Sept. 2024
ML and Data Intern — *Fremont, CA*

- Developed a **Langchain**-based product tool for Walmart sellers reducing manual review time by 25%.
- Enabled data access for sales teams by implementing a Text-to-SQL AI Agent using **Langgraph**.
- Generated 2400+ qualified leads through customer data analysis using **Pandas** and **PowerBI**.
- Automated **MySQL** database management, reducing query processing time by 75%.

CodeLab - Tech Consulting Club — Oct. 2022 – Jun. 2023
Product Manager & Events Lead — *Ann Arbor, MI*

- Led a 9-person team to build and deploy an E-commerce product image generator with **FastAPI** and **DALL·E**, reducing seller costs by 17%.

Open Source Projects

Tailored Academic & Resource Assistant | *Python, RAG, Streamlit, VectorDB* — July 2024 – Present

- Built an **agentic chatbot** serving 150+ students for course selection and career guidance.
- Reduced response latency by 40% using optimized prompt engineering, caching, and ensemble models.
- Created documentation enabling 10+ student projects to implement similar **RAG**-based systems.

IOLBench- LLM Benchmarking | *ACL Pending* — May 2024 – Oct. 2024

- Constructed a dataset of **1,000+ linguistic problems**, enabling industry-wide model benchmarking.
- Measured performance of 8 LLMs across language and time-series tasks using an evaluation pipeline.
- Created a **Python library** with **NumPy** and **RegEx** for scaling model evaluations of LLMs.

Technical Skills

Languages (Proficient): Python, SQL
Languages (Familiar): Java, C++, JavaScript, Go
ML/AI/Data: PyTorch, RAG, Scikit-Learn, Agentic AI, Transformers, PostgreSQL, Pandas
Cloud & DevOps: vLLM, LM Cache, AWS (EC2, Lambda, SES), Docker, Kubernetes, FastAPI