

Yash Mahajan, Ph.D.

Natural Language Processing | LLM Evaluation | Reliable and Efficient AI Systems

yash.mahajan50@gmail.com +1 334-467-6545 Orlando, FL, USA

linkedin.com/in/yash-mahajan github.com/yzm1205

Research Profile

I study how language models and AI systems should be evaluated, selected, and deployed when aggregate benchmark scores do not fully capture semantic behavior, user requirements, or deployment constraints. My work combines natural language processing, model evaluation, scalable inference, and agentic systems to develop reliable and reproducible methods for matching AI models to real-world tasks.

Research interests: reliable capability and alignment evaluation; resource-aware model selection and routing; evaluation of multi-step and agentic AI systems; efficient multi-GPU inference; sentence representation and semantic alignment.

Academic Appointments

University of Central Florida

Postdoctoral Researcher / AI Engineer, BridgeAI Lab

Dec. 2025–Present

Orlando, Florida

- Lead the design of **LLMCompass**, a research platform for studying capability-aware and resource-aware selection among more than 50 language models across more than 25 task skills.
- Develop research protocols for evaluating instruction adherence, response quality, semantic alignment, ranking stability, and task–model fit across multiple datasets and model families.
- Build reproducible multi-model and multi-dataset evaluation pipelines using PyTorch, Hugging Face, vLLM, FastAPI, PostgreSQL, Docker, Slurm, and multi-GPU infrastructure.
- Implement scalable inference and model-sharding workflows for large-model experimentation on H100 and NSF-supported computing resources.
- Design modular experiment infrastructure that records model, prompt, dataset, hardware, latency, throughput, and quality metadata for controlled comparative studies.
- Investigate automated mappings from user intent to capability requirements, model rankings, and deployment configurations.
- Collaborate on research using topological data analysis to study and optimize model training behavior.

Education

Auburn University

Ph.D., Computer Science and Software Engineering; GPA: 3.77/4.00

Dec. 2025

Auburn, Alabama

Dissertation: *Evaluating Semantic and Contextual Alignment in Language Models*

Advisor: Sathyanarayanan N. Aakur **Co-advisor:** Shubhra Kanti Karmaker

Auburn University

M.S., Computer Science and Software Engineering

Auburn, Alabama

Rajiv Gandhi Proudhyogiki Vishwavidyalaya

B.Tech., Computer Science and Engineering; GPA: 3.7/4.0

2019

India

Peer-Reviewed Publications

1. **Yash Mahajan**, Matthew Freestone, Naman Bansal, Sathyanarayanan N. Aakur, and Santu Karmaker. “Revisiting Word Embeddings in the LLM Era.” In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 2686–2717, 2025. DOI: 10.18653/v1/2025.ijcnlp-long.145.
2. **Yash Mahajan**, Naman Bansal, Eduardo Blanco, and Santu Karmaker. “ALIGN-SIM: A Task-Free Test Bed for Evaluating and Interpreting Sentence Embeddings through Semantic Similarity Alignment.” In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7393–7428, 2024. DOI: 10.18653/v1/2024.findings-emnlp.436.
3. S. Sarkar, B. S. Bijoy, S. J. Saba, D. Feng, **Yash Mahajan**, M. R. Amin, S. R. Islam, and S. K. Karmaker. “Ad-hoc Monitoring of COVID-19 Global Research Trends for Well-Informed Policy Making.” *ACM Transactions on Intelligent Systems and Technology*, 14(2), pp. 1–28, 2023.

Current Research Program and Software

LLMCompass: Evaluation and Resource-Aware Model Selection

UCF, ongoing

- Research question: how can an AI system select a model that satisfies task-specific capability requirements while jointly accounting for quality, uncertainty, latency, memory, throughput, and monetary or computational cost?
- Curate skill-oriented evaluations for summarization, question answering, reasoning, instruction following, and related capabilities; compare rankings across datasets, prompts, judges, and deployment settings.
- Study generalization of model rankings to unseen tasks and evaluate whether intent-to-skill mappings provide reliable evidence for model recommendation.
- Develop reproducible inference and data infrastructure as a byproduct of the research, including model execution, metadata capture, ranking, and deployment-plan generation.

ALIGN-SIM Resources

Auburn University

- Released software and data for task-free evaluation of sentence embeddings using five semantic similarity alignment criteria.
- Evaluated 13 classical and LLM-induced encoders and demonstrated that strong downstream benchmark performance can coexist with failures on intuitive semantic alignment tests.

Shard-Any-LLMs

Selected research software

- Developed an automated model-sharding workflow for distributing large language models across multiple GPUs, with command-line interfaces, Hugging Face integration, and environment-based authentication.

Research Funding, Computing, and Proposal Experience

- **NSF-supported GPU compute allocation:** received access to GPU resources valued at approximately \$600,000 for large-scale language-model evaluation and inference research.
- Contributed to the development of multiple faculty-led grant proposals through technical writing, problem formulation, methodology design, preliminary experimentation, and research-infrastructure planning.

Teaching Experience

Auburn University

Aug. 2020–Dec. 2025

Graduate Teaching Assistant, Computer Science and Software Engineering

Auburn, Alabama

- Graded coursework and examinations and held regular office hours across multiple undergraduate and graduate Computer Science courses.
- Designed assessments from the ground up for **Game Development** and **Quantum Computing**; developed grading criteria, evaluated student work, and supported students through office hours.
- Led **Python** and **MATLAB** laboratory sections of approximately 50 students three times per week for four semesters.

- Led **Data Structures and Algorithms** laboratory sections of approximately 50 students for two semesters.
- Delivered three class lectures in a graduate-level **Machine Learning** course in Fall 2020; graded assignments and held office hours.

Student Mentoring and Academic Leadership

University of Central Florida

Undergraduate Research Mentor, Undergraduate Research Experience (URE) Program

Spring 2025

Orlando, Florida

- Mentored nine undergraduate researchers in one to two meetings per week.
- Introduced foundational concepts in natural language processing and guided students in understanding practical NLP applications and research workflows.

Auburn University

Volunteer AI Team Leader

During doctoral study

Auburn, Alabama

- Led weekly meetings for undergraduate students, introduced AI research topics, encouraged participation in projects, and supported students developing early research interests.

Guest Lectures

- Delivered three guest lectures to master's students at the University of Central Florida, with approximately 30 students per session, across courses in Machine Learning and transportation-related programs.

Professional and University Service

- Reviewer, **ACL Rolling Review (ARR)**.
- Reviewer, **ACM Transactions on Internet Technology (TOIT)**.
- Judge, university student summits and project presentations on multiple occasions.

Prior Industry and Research Experience

LogikView

Data Analyst Intern

Jul. 2019–Dec. 2019

Indore, India

- Built interactive dashboards and improved data-processing workflows for customer relationship management analytics.

Technex / IIT BHU Program

Machine Learning Intern

May 2019–Jul. 2019

India

- Developed and evaluated machine-learning models for disease-risk prediction using hospital medical-record data.
- Built dashboards to support clinical users in reviewing patient histories and trends.

Technical Expertise

Machine learning and NLP: PyTorch, Hugging Face Transformers, sentence embeddings, language-model evaluation, benchmarking, metrics design, alignment analysis.

LLM systems: vLLM, text-generation inference, multi-GPU inference, model sharding, agentic workflows, LangChain, MCP servers.

Software and infrastructure: Python, Bash, FastAPI, Docker, Linux, REST APIs, PostgreSQL/pgvector, Slurm, Git, Weights & Biases, Streamlit, Pandas, NumPy.

Selected Courses Prepared to Teach

Artificial Intelligence; Machine Learning; Natural Language Processing; Data Structures and Algorithms; Python Programming; Applied Deep Learning; Large Language Model Systems; Evaluation and Responsible Deployment of Generative AI; AI Agents and Tool-Using Systems.