

Pratik Shah

[✉ pshah479@gatech.edu](mailto:pshah479@gatech.edu) | [📞 +1 \(404\)-428-6340](tel:+1(404)4286340) | [LinkedIn](https://www.linkedin.com/in/shahpratik02) | [GitHub](https://shahpratik02.github.io)

Education

Georgia Institute of Technology , Atlanta, USA <i>M.S. in Computer Science (Specialization: Machine Learning)</i>	[Aug '24 - May '26] Overall GPA: 3.87/4.00
Indian Institute of Technology (IIT) Bombay , Mumbai, India <i>B.Tech in Mechanical Engineering with Honors Department Rank 5 among 190+ students</i> <i>Minor in Data Science and Artificial Intelligence</i>	[Nov '20 - May '24] Overall GPA: 9.44/10.00

Publications

- "RANGER: Repository-Scale Agent for Graph-Enhanced Retrieval" Submitted in **ICLR 2026** [arXiv:2509.25257](https://arxiv.org/abs/2509.25257)
"Lagrangian Index Policy for Restless Bandits with Average Reward" Submitted in **Queueing Systems Journal** [arXiv:2412.12641](https://arxiv.org/abs/2412.12641)
"Reinforcement Learning in non-Markovian Environments" Published in **Systems and Control Letters Journal** vol. 185, 105751

Work Experience

Nutanix Intern, Member of Technical Staff	[May '25 - Aug '25]
• Developed RANGER a repository-scale agent utilising RL-enhanced GraphRAG for code tasks Provisional Patent & ICLR '26	
• Created a Monte Carlo Tree Search (MCTS) based graph retrieval algorithm fusing bi-encoder speed with cross-encoder precision	
• Built an AST -based tool to construct Neo4j knowledge graphs of entire repos, capturing hierarchical and cross-file dependencies	
• Developed a dual-stage retriever combining text2cypher for entity lookup with the novel MCTS algorithm for graph traversal	
• Beat Qwen-3-8B (SOTA) semantic retrieval, scoring 6% higher NDCG@10 on CodeSearchNet (NL→Code benchmark). Got 6% higher exact match on CrossCodeEval and 5% higher accuracy on RepoBench for code completion and retrieval over baselines	
Partnership for an Advanced Computing Environment (PACE) Graduate Research Assistant	[Jan '25 - Present]
• Working on AI inference server with a LiteLLM gateway routing requests to vLLM servers, scheduled on HPC GPUs via slurm	
• Enabled 51 courses to use PACE's HPC clusters by containerizing ML workloads, configuring shared storage and scheduling jobs	
• Developed workshops for the AI Makerspace , a university-wide initiative with Nvidia for hands-on AI/ML education, covering multi-GPU training (torchrun), Llama-2 fine-tuning, and model deployment with TensorRT and Triton Inference Server	
Microsoft Data Science Intern	[May '23 - Jun '23]
• Automated personalized health tips generation using OpenAI GPT Models on MSN health pages data In Production	
• Implemented an automated RAG pipeline from scratch using serverless Azure Functions , created REST APIs to retrieve contextual data from Azure SQL , and leveraged the OpenAI Completions API to interact with GPT-3.5 for generating tips	
• Reduced the tip generation time from 2 weeks to 30 minutes for 100 tips and attained a per-tip cost of ~ \$0.0015	
• Created a GPT-3.5 based translation pipeline, expanding coverage from 14 English to all 24 markets, including non-English ones	
Data Axe Data Science Intern	[May '22 - Jul '22]
• Consolidated 50,000 job titles into 1,000 standardized titles using NLP and clustering for the company's lead generation service	
• Applied tokenization, GloVe vectorization, dimensionality reduction (PCA , t-SNE), and K-means clustering to group job titles	

Research Projects

Lagrangian Index Policy (LIP) for Restless Bandits With Average Reward	[Jul '23 - Dec '24]
• Designed an index policy for restless bandits to optimize long-run rewards, with applications in resource allocation and scheduling	
• LIP requires no indexability conditions and the proposed tabular and NN-based reinforcement learning schemes for model-free setting require significantly less memory and time than the Whittle Index Policy (WIP), which is the standard in this domain	
• The new policy is asymptotically optimal and applicable to both Whittle Indexable and Non-Whittle Indexable problems	
Reinforcement Learning in Non-Markovian Environments	[Dec '22 - Sep '23]
• Designed a new RL agent, the Non-Markovian Q Agent (NMQ), to tackle environments where past information is crucial	
• The NMQ agent uses an autoencoder -based scheme to tackle non-Markovianity by learning a latent state space for a Deep Q-Network (DQN). Modified OpenAI Gym environments like CartPole to be partially observable for testing the agent	
• The NMQ agent outperformed the standard DQN agent in partially observable environments and Non-Markovian random walks	

Technical Skills

Skills	AWS Certified Cloud Practitioner , Python, C++, SQL, Azure, Spark, Java, CUDA, Linux, Neo4j, Git, Slurm
Frameworks	PyTorch, TensorFlow, vllm, LangChain, LlamaIndex, HuggingFace, OpenAI, Gym, RLlib, torchrun, TesnorRT

Extracurricular Activities and Awards

- **Scholarships:** [KCMET Fellowship](#) ['24], [NFIA Scholarship](#) ['24], [KVPY Fellowship](#) ['19 & '20]
• Led IITB's [autonomous underwater vehicle team](#) on [L&T](#) Defence ROV and [ONGC](#) subsea inspection project [Aug '22 - May '23]
• Elected as a student **mentor** for **14** freshmen and **4** sophomores, offering academic and general guidance [May '22 - May '24]