Pratik Shah

pratik2002shah@gmail.com | +1 (404)-428-6340 | www.linkedin.com/in/shahpratik02 | https://shahpratik02.github.io

Education

Georgia Institute of Technology, Atlanta, USA [Aug '24 - May '26]
M.S. in Computer Science (Specialization: Machine Learning) Overall GPA: **3.9/4.0**

Indian Institute of Technology Bombay, Mumbai, India [Nov '20 - May '24]
*B.Tech in Mechanical Engineering with Honors | Department Rank **5** among 190+ students* Overall GPA: **9.44/10.00**
Minor in Data Science and Artificial Intelligence

Publications

"RANGER: Repository-Scale Agent for Graph-Enhanced Retrieval" Under Review in **ICLR 2026** *arXiv:2509.25257*

"Lagrangian Index Policy for Restless Bandits with Average Reward" Submitted in **Queueing Systems** Journal *arXiv:2412.12641*

"Reinforcement Learning in non-Markovian Environments" Published in **Systems & Control Letters** Journal vol. 185, 105751

Work Experience

Georgia Tech HPC Center (PACE) | Graduate Research Assistant [Jan '25 - Present]
- Developing a multimodal AI inference server running GPT-OSS-120B, InternVL-3.5, and SD-XL, using vLLM for LLM/VLM tasks and TensorRT-engine Triton server for image generation, unified by a custom OpenAI-style API wrapper for LiteLLM integration
- Orchestrated LiteLLM request routing across ephemeral Slurm-scheduled GPU nodes running Apptainer-based inference servers, and built a cron-driven self-healing system for automated service discovery and failover in a non-Kubernetes HPC environment
- Working on the AI Makerspace, a campus initiative, leading tutorials like implementing FlashAttention in CUDA from scratch

Nutanix | Intern, Member of Technical Staff [May '25 - Aug '25]
- Developed RANGER a repository-scale agent utilizing 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

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

Key Projects

FlashAttention-2 CUDA Kernels & LLM Inference (Code) [Sep '25 - Dec '25]
- Implemented FlashAttention-2 CUDA kernels using block-tiling and online softmax; **2.59** faster than PyTorch on H100
- Optimized autoregressive decoding with KV-cache update & FlashAttention-2 decode kernel, achieving **3.1×** lower TBT latency
- Built an end-to-end LLM inference engine with RoPE embeddings, lazy KV initialization and dedicated prefill/decode kernels

Lagrangian Index Policy (LIP) for Restless Bandits With Average Reward [Jul '23 - Dec '24]
- Developed LIP, a Deep RL policy, for Restless Bandits backed by theoretical optimality proofs, outperforming the SOTA Whittle Index Policy in both memory efficiency and long-run rewards for constrained resource allocation and scheduling tasks

Reinforcement Learning (RL) in Non-Markovian Environments [Dec '22 - Sep '23]
- Developed a Recurrent State-Space Model (RSSM) in TensorFlow to solve partially observed control tasks, utilizing a recursive predictive autoencoder to learn latent dynamics from historical data which are then used by a Deep Q-Network to take actions
- Built an alternating RSSM training pipeline with hybrid MSE + BCE losses that improved episodic rewards by **5×** over baseline

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, TensorRT

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]