Shreyas Pimpalgaonkar

Website: shreyaspimpalgaonkar.github.io • Email: shreyas.pgkr@gmail.com • Phone: +1 (332) 254-8873

Education

Courant Institute of Mathematical Sciences, New York University — *Sep 2022 - May 2024*
Master of Science in Computer Science — GPA: 4.0/4.0

Indian Institute of Technology Bombay — *Jul 2016 - May 2020*
Bachelor of Science in Computer Science with Honors — GPA: 8.82/10.0

Experience

Citadel | *Quantitative Research Intern* — *Jun 2023 - Aug 2023*

- Designed and implemented alpha capture signals for Convertible Bonds strategies, leading to a 2.48% quarterly backtested profit
- Modeled realized volatility measures to evaluate trader's implied vol marks and implemented an automated feedback framework

GRAIL Lab, NYU | *Deep Learning Researcher, Guide: Prof. Lerrel Pinto* — *Jan 2023 - present*

- Creating robot foundation world models to execute 100+ tasks and improve the efficiency of robot learning in real-world scenarios
- Implemented BC, RL, MPC policies and world models for locomotion and manipulation tasks in 5 robot simulation environments
- Proposed changes to architectures and training algorithms to improve a robot's new task learning speed by up to 50%

CILVR Lab, NYU | *Deep Learning Researcher, Guide: Prof. He He* — *Jan 2024 - present*

- Improved the arithmetic and mathematical reasoning capabilities of large language models using novel contrastive approaches
- Experimenting with olympiad geometry figure extraction using VLMs and generative processes for grad level MATH QA datasets
- Exploring mechanistic interpretability approches for enabling LLM interpolation to models of variable sizes and architectures

Goldman Sachs | *Quantitative Researcher and Developer, Intern + Full-time* — *Summer '19, Jul 2020 - Aug 2022*

- Implemented capital optimization frameworks totalling to $10Bn RWA in accordance with new FRTB-CVA and SACCR regulations
- Executed 100+ modeling and infrastructure enhancements, including software engineering, quantitative model updates, new regulation modeling, API design, calculation and process speedups, database and cache optimizations, and infrastructure scaling
- Enhanced Machine Learning models to attribute PnL tail movements to assets, achieving 90% speedup and 70% higher accuracy
- Obtained the highest performance rating (exceeding expectations) across all performance metrics

CFILT Lab, IIT Bombay | *NLP Researcher, Prof. Pushpak Bhattacharyya* — *Jan 2020 - Aug 2020*

- Innovated Aspect Based Emotion Analysis using BERT based models to identify 6 common emotions in restaurant reviews
- Released datasets containing 3,000+ emotion-aspect pairs, and achieved 91% accuracy and a 0.83 F1 score on test set
- Published a research paper in ICON 2021 titled ProverbNet, a multilingual database containing 3400+ proverbs and metadata

Ubisoft | *Deep Learning Research Intern, Autonomous Driving Vehicles* — *May 2018 - Jul 2018*

- Engineered an in-game self driving AI using perception and navigation techniques that achieved an 80% OOD success rate

Key Projects

- **RAG Search Engine**: Engineered a custom search engine utilizing a fine-tuned Mistral-Instruct-7b model with queries augmented using relevant context generated by a search API and FAISS, resulting in improved factual correctness of the model.
- **Machine Unlearning for LLMs**: Enhanced a method for unlearning concepts learned by LLMs and applied it to Llama2-7b to unlearn Harry Potter, resulting in improved pipeline and under 2% impact on the model's performance on popular eval benchmarks.
- **Model Based Imitation Learning**: Produced a novel model based imitation learning framework utilizing dreamer-v3 to run expert guided policies using OT based rewards on a simple walker-walk task to achieve 800+ reward in 15k env steps.
- **SoundBox**: Designed an RL agent using Policy Gradient algorithms to act in an environment of a realistic human vocal tract and generate English letter sounds. Working on generating words, sentences, and a text to speech agent.
- **Distributed Ring Attention**: Implemented Flash Attention v2 and Ring Attention from scratch using Pytorch, CUDA and Triton. Achieved 64k context length on 4 RTX8000 GPUs on Gemma 2b model, an improvement of 4x compared to vanilla attention.
- **Bayesian Optimization**: Investigated Bayesian Optimization (BO) as an alternative to MLE for optimizing complex likelihood functions and created a hybrid approach 2.5x faster than MLE and a solution with 3x likelihood compared to BO.
- **Seminar Research and Presentation**: Reviewed 20+ and presented 4 recent advances in foundational models, diffusion models, DL infrastructure, DL theory (optimization, approximation, generalization) as part of a semester long research seminar.

Others

- **Skills**: C++, Python, SQL, LLMs, VLMs, RL, Pytorch, JAX, Huggingface, Langchain, Robotics, Mujoco, Gym, AWS, HPC
- **NYC Generative AI Hackathon**: Secured first place for presenting a novel text to interactive comic app using GPT + DALLE
- **Competitive Programming**: Secured 36th position in ACM ICPC GNYR 2022; held a maximum codechef rating of 1928
- **Teaching Assistant**: Fundamental Algorithms, Convex Optimization, Portfolio Management graduate level courses
- **JEE Entrance**: Secured All India 99.95 percentile in IIT JEE Main 2016 and and 99.51 percentile in IIT JEE Advanced 2016
- **Miscellaneous**: Elected as the department social secretary during undergrad; part of the music club and running club